Adaptive density estimation: a curse of support?

Patricia Reynaud-Bouret, Vincent Rivoirard, Christine Tuleau-Malot

Adaptive density estimation: a curse of support?☆

Patricia REYNAUD-BOURET[*,a], Vincent RIVOIRARD[b,c], Christine TULEAU-MALOT[a]

[a] *Laboratoire Jean-Alexandre Dieudonné, CNRS UMR 6621, Université de Nice Sophia-Antipolis, Parc Valrose, 06108 Nice Cedex 2, France.*
[b] *Laboratoire de Mathématiques, CNRS 8628, Université de Paris Sud, 91405 Orsay Cedex, France.*
[c] *Département de Mathématiques et Applications, ENS-Paris, 45 rue d'Ulm, 75230 Paris Cedex 05, France.*

Abstract

This paper deals with the classical problem of density estimation on the real line. Most of the existing papers devoted to minimax properties assume that the support of the underlying density is bounded and known. But this assumption may be very difficult to handle in practice. In this work, we show that, exactly as a curse of dimensionality exists when the data lie in $\mathbb{R}^d$, there exists a curse of support as well when the support of the density is infinite. As for the dimensionality problem where the rates of convergence deteriorate when the dimension grows, the minimax rates of convergence may deteriorate as well when the support becomes infinite. This problem is not purely theoretical since the simulations show that the support-dependent methods are really affected in practice by the size of the density support, or by the weight of the density tail. We propose a method based on a biorthogonal wavelet thresholding rule that is adaptive with respect to the nature of the support and the regularity of the signal, but that is also robust in practice to this curse of support. The threshold, that is proposed here, is very accurately calibrated so that the gap between optimal theoretical and practical tuning parameters is almost filled.

Key words: Density estimation, Wavelet, Thresholding rule, infinite support
2000 MSC: 62G05, 62G07, 62G20

1. Introduction

This paper deals with the classical problem of density estimation for unidimensional data. Our aim is to provide an adaptive method which requires as few assumptions as possible on the underlying density in order to apply it in an exploratory way. In particular, we do not want to have any assumption on the density support. Moreover this method should be quite easy to implement and should have good theoretical performance as well.

Density estimation is a task that lies at the core of many data preprocessing. From this point of view, no assumption should be made on the underlying function to estimate. At least in a first approach, histograms or kernel methods are often used. These popular linear estimators do not require any assumption on the support and have good theoretical performance. The main problem is to choose the bandwidth, which is usually performed by cross-validation (see the fundamental paper by Rudemo (1982)) or by other data-driven methods (see Silverman (1986)). However, most of the time, the bandwidth is selected uniformly in

☆The authors acknowledge the support of the French Agence Nationale de la Recherche (ANR), under grant ATLAS (JCJC06_137446) "From Applications to Theory in Learning and Adaptive Statistics".

*Corresponding author. Tel.: (+33) 04 92 07 60 33; Fax: (+33) 04 93 51 79 74.
Email addresses: reynaudb@unice.fr (Patricia REYNAUD-BOURET), Vincent.Rivoirard@math.u-psud.fr (Vincent RIVOIRARD), malot@unice.fr (Christine TULEAU-MALOT)

space, which suffers several drawbacks due to the lack of spatial adaptivity. To improve this point, Sain and Scott (1996) have proposed a kernel procedure which makes the choice of the bandwidth more local, this procedure being still based on intensive cross-validation. It is worth noting that these kernel methods may have a high computational cost, are often based on asymptotic arguments and do not provide theoretical guarantees from the adaptive minimax point of view.

One possible adaptive minimax approach is to consider model selection. Following Akaike's criterion for histograms, Castellan (2000) has derived adaptive minimax procedures for density estimation (see Massart (2007) for detailed proofs and Birgé and Rozenholc (2006) for a practical point of view). To remedy the lack of smoothness of histograms, piecewise polynomial estimates can also be used (see for instance Castellan (2003), Willett and Nowak (2007) or Koo *et al.* (1999) for the spline basis). It is worth emphasizing that, basically, the necessary input of model selection methods is the support of the underlying density, classically assumed to be $[0, 1]$. In practice, the data are usually rescaled by the smallest and largest observations before performing any of the previous algorithms. This preprocessing has not been studied theoretically. In particular, what happens if the density is heavy-tailed? Note that ℓ_1-penalty methodologies can also be used, providing oracle inequalities without any support assumption (see for instance Bertin *et al.* (2010)), but minimax properties have not been investigated for such estimators.

Now let us turn to wavelet thresholding. Donoho *et al.* (1996) have first provided theoretical adaptive minimax results in the density setting. This paper is a theoretical benchmark but their threshold depends on the extraknowledge of the infinite norm of the underlying density. In practice, even if this quantity is known, this choice is often too conservative. From a computational point of view, the DWT algorithm combined with a keep or kill rule on each coefficient makes these methods as one of the easiest adaptive methods to implement, once the threshold is known. Here lies the fundamental problem: after rescaling and binning the data as in Antoniadis *et al.* (1999) for instance, one can reasonably think that the number of observations in a "not too small" interval is Gaussian, up to some eventual transformation. So basically the thresholding rules adapted to the Gaussian regression setting should work here even if many assumptions are required. Even if in Brown *et al.* (2010) theoretical justifications are given, the method still relies heavily on the precise knowledge of the support which is directly linked to the size of the bins. In their seminal work Herrick *et al.* (2001) have already observed that in practice the basic Gaussian approximation for general wavelet bases is quite poor. This can be corrected by the use of the Haar basis and accurate thresholding rules but the reconstructions are consequently piecewise constant. Note also that in this paper no assumption was made on the support of the underlying density. More recently, Juditsky and Lambert-Lacroix (2004) have proposed an adaptive thresholding procedure on the whole real line. Their threshold is not based on a direct Gaussian approximation. Indeed, the chosen threshold depends randomly on the localization in time and frequency of the coefficient that has to be kept or killed. They derive adaptive minimax results for Hölderian spaces, exhibiting rates that are different from the bounded support case. However there is a gap between their optimal theoretical and practical tuning parameters of the threshold.

If the main goal of this paper is to investigate assumption-free wavelet thresholding methodologies as explained in the first paragraph, we also aim at fulfilling this gap by designing a new threshold depending on a tuning parameter γ: the precise form of the threshold is closely related to sharp exponential inequalities for iid variables, avoiding the use of Gaussian approximation. Unlike methods of Juditsky and Lambert-Lacroix (2004) and Herrick *et al.* (2001), all the coefficients (and in particular the coarsest ones) are likely to be thresholded. Moreover, since our threshold is defined very accurately from a non asymptotic point of view, we obtain sharp oracle inequalities for $\gamma > 1$. But we also prove that taking $\gamma < 1$ deteriorates the theoretical properties of our estimator. Hence the remaining gap between theoretical and practical thresholds lies in a second order term (see Section 2 for more details). The construction of our estimators and the previous results are stated in Section 2. Next, in Section 3, we illustrate the impact of the bounded support assumption by exhibiting minimax rates of convergence on the whole class of Besov spaces extending for the $\mathbb{L}_2$ loss the results of Juditsky and Lambert-Lacroix (2004). In particular, when the support is infinite, our results reveal how minimax rates deteriorate according to the sparsity of the density. We also show that our estimator is adaptive minimax (up to a logarithmic term) over Besov balls with respect to the regularity but also with respect to the support (finite or not). In Section 4, we investigate the curse of support for the most well-known support-dependent methods and compare them with our method and with the cross-validated kernel method. Our method, which is naturally spatially adaptive, seems to be robust with respect to the size of the support or the tail of the underlying density. We also implement our method on real data, revealing the potential impact of our methodology for practitioners. The appendices are dedicated to an analytical description of the biorthogonal wavelet basis but also to the proofs of the main results.

2. Our method

Let us observe a n-sample of density f assumed to be in $\mathbb{L}_2(\mathbb{R})$. We denote this sample $X_1, \dots, X_n$. We estimate f via its coefficients on a special biorthogonal wavelet basis, due to Cohen *et al.* (1992). The decomposition of f on such a basis takes the following form:

$$f = \sum_{k \in \mathbb{Z}} \beta_{-1k} \tilde{\psi}_{-1k} + \sum_{j \geq 0} \sum_{k \in \mathbb{Z}} \beta_{jk} \tilde{\psi}_{jk}, \tag{2.1}$$

where for any $j \geq 0$ and any $k \in \mathbb{Z}$,

$$\beta_{-1k} = \int_{\mathbb{R}} f(x) \psi_{-1k}(x) dx, \quad \beta_{jk} = \int_{\mathbb{R}} f(x) \psi_{jk}(x) dx.$$

The most basic example of biorthogonal wavelet basis is the Haar basis where the father wavelets are given by

$$\forall k \in \mathbb{Z}, \quad \psi_{-1k} = \tilde{\psi}_{-1k} = \mathbb{1}_{[k;k+1]}$$

and the mother wavelets are given by

$$\forall j \geq 0, \ \forall k \in \mathbb{Z}, \quad \psi_{jk} = \tilde{\psi}_{jk} = 2^{j/2} \left(\mathbb{1}_{[k2^{-j};(k+1/2)2^{-j})} - \mathbb{1}_{[(k+1/2)2^{-j};(k+1)2^{-j}]} \right).$$

The other examples we consider are more precisely described in Appendix A. The essential feature is that it is possible to use, on the one hand, decomposition wavelets ψ_{jk} that are piecewise constants, and, on the other hand, smooth reconstruction wavelets $\tilde{\psi}_{jk}$. In particular, except for the Haar basis, decomposition and reconstruction wavelets are different. To shorten mathematical expressions, we set

$$\Lambda = \{(j,k) : \quad j \geq -1, k \in \mathbb{Z}\} \tag{2.2}$$

and (2.1) can be rewritten as

$$f = \sum_{(j,k) \in \Lambda} \beta_{jk} \tilde{\psi}_{jk} \quad \text{with} \quad \beta_{jk} = \int \psi_{jk}(x) f(x) dx. \tag{2.3}$$

A classical unbiased estimator for β_{jk} is the empirical coefficient

$$\hat{\beta}_{jk} = \frac{1}{n} \sum_{i=1}^{n} \psi_{jk}(X_i), \tag{2.4}$$

whose variance is σ_{jk}^2/n where

$$\sigma_{jk}^2 = \int \psi_{jk}^2(x) f(x)dx - \left(\int \psi_{jk}(x) f(x)dx\right)^2.$$

Note that σ_{jk}^2 is classically unbiasedly estimated by $\widehat{\sigma}_{jk}^2$ with

$$\widehat{\sigma}_{jk}^2 = \frac{1}{n(n-1)} \sum_{i=2}^{n} \sum_{l=1}^{i-1} (\psi_{jk}(X_i) - \psi_{jk}(X_l))^2.$$

Now, let us define our thresholding estimate of f. In the sequel there are two different kinds of steps, depending on whether the estimate is used for theoretical or practical purposes. Both situations are respectively denoted '*Th.*' and '*Prac*'.

Step 0

Th. Choose a constant $c \geq 1$, a real number c' and let j_0 such that $j_0 = \lfloor \log_2([n^c(\log n)^{c'}]) \rfloor$. Choose also a positive constant γ.

Prac. Let $j_0 = \lfloor \log_2(n) \rfloor$.

Step 1 Set $\Gamma_n = \{(j,k) : -1 \leq j \leq j_0,\ k \in \mathbb{Z}\}$ and compute for any $(j,k) \in \Gamma_n$, the non-zero empirical coefficients $\hat{\beta}_{jk}$ (whose number is almost surely finite).

Step 2 Threshold the coefficients by setting $\tilde{\beta}_{jk} = \hat{\beta}_{jk} \mathbb{1}_{|\hat{\beta}_{jk}| \geq \eta_{jk}}$ according to the following threshold choice.

Th. Overestimate slightly the variance σ_{jk}^2 by using

$$\widetilde{\sigma}_{jk}^2 = \widehat{\sigma}_{jk}^2 + 2\|\psi_{jk}\|_\infty \sqrt{2\gamma \widehat{\sigma}_{jk}^2 \frac{\log n}{n}} + 8\gamma \|\psi_{jk}\|_\infty^2 \frac{\log n}{n}$$

and choose

$$\eta_{jk} = \eta_{jk,\gamma} = \sqrt{2\gamma \widetilde{\sigma}_{jk}^2 \frac{\log n}{n}} + \frac{2\|\psi_{jk}\|_\infty \gamma \log n}{3n}. \tag{2.5}$$

Prac. Estimate unbiasedly the variance by $\widehat{\sigma}_{jk}^2$ and choose

$$\eta_{jk} = \eta_{jk}^{Prac} = \sqrt{2\widehat{\sigma}_{jk}^2 \frac{\log n}{n}} + \frac{2\|\psi_{jk}\|_\infty \log n}{3n}. \tag{2.6}$$

Step 3 Reconstruct the function by using the $\tilde{\beta}_{jk}$'s and denote

Th.

$$\tilde{f}_{n,\gamma} = \sum_{(j,k)\in\Gamma_n} \tilde{\beta}_{jk} \tilde{\psi}_{jk} \tag{2.7}$$

Prac.

$$\tilde{f}_n^{Prac} = \left(\sum_{(j,k)\in\Gamma_n} \tilde{\beta}_{jk} \tilde{\psi}_{jk} \right)_+ \tag{2.8}$$

Note that this method can easily be implemented with a low computational cost. In particular, unlike the DWT-based algorithms, our algorithm does not need numerical approximations, except at **Step 3** for the computations of the $\tilde{\psi}_{jk}$ (unless, we use the Haar basis). However, a preprocessing, independent of the algorithm, can be used to compute reconstruction wavelets at any required precision. Both practical and theoretical thresholds are based on the following heuristics. Let $c_0 > 0$. Define the heavy mass zone as the set of indices $(j,k) \in \Lambda$ such that $f(x) \geq c_0$ for x in the support of ψ_{jk} and $\|\psi_{jk}\|_\infty^2 = o_n(n(\log n)^{-1})$. In this heavy mass zone, the random term of (2.5) or (2.6) is the main one and we asymptotically derive that with large probability

$$\eta_{jk,\gamma} \approx \sqrt{2\gamma\widetilde{\sigma}_{jk}^2 \frac{\log n}{n}} \quad \text{and} \quad \eta_{jk}^{Prac} \approx \sqrt{2\widehat{\sigma}_{jk}^2 \frac{\log n}{n}}. \tag{2.9}$$

The shape of the right hand terms in (2.9) is classical in the density estimation framework (see Donoho *et al.* (1996)). In fact, they look like the threshold proposed by Juditsky and Lambert-Lacroix (2004) or the universal threshold η^U proposed by Donoho and Johnstone (1994) in the Gaussian regression framework. Indeed, we recall that, in this set-up,

$$\eta^U = \sqrt{2\sigma^2 \log n},$$

where σ^2 (assumed to be known in the Gaussian framework) is the variance of each noisy wavelet coefficient. Actually, the deterministic term of (2.5) (or (2.6)) constitutes the main difference with the threshold proposed by Juditsky and Lambert-Lacroix (2004): it replaces the second keep or kill rule applied by Juditsky and Lambert-Lacroix on the empirical coefficients. This additional term allows to control large deviation terms for high resolution levels. It is directly linked to Bernstein's inequality (see the proofs in Appendix B). The forthcoming oracle inequality (Theorem 1) holds with (2.5) for any $\gamma > 1$: this is essential to fulfill the gap between theory and practice. Indeed, note that if one takes $c = \gamma = 1$ and $c' = 0$ then the main difference between (2.5) and (2.6) is a second order term in the estimation of σ_{jk}^2 by $\widetilde{\sigma}_{jk}^2$. But the main part is exactly the same: when the coefficient lies in the heavy mass zone and when γ tends to 1, $\eta_{jk,\gamma}$ tends to η_{jk}^{Prac} with high probability. Indeed, one can note that for all $\varepsilon > 0$ and $\gamma > 1$,

$$\eta_{jk}^{Prac} \leq \eta_{jk,\gamma} \leq \sqrt{2\gamma(1+\varepsilon)\widehat{\sigma}_{jk}^2 \frac{\log n}{n}} + \left(\frac{2}{3} + \sqrt{2(8+2\varepsilon^{-1})}\right) \frac{\|\psi_{jk}\|_\infty \gamma \log n}{n}.$$

As often suggested in the literature, instead of estimating $\mathrm{Var}(\hat{\beta}_{jk})$, we could have used the inequality

$$\mathrm{Var}(\hat{\beta}_{jk}) = \frac{\sigma_{jk}^2}{n} \leq \frac{\|f\|_\infty}{n}$$

and we could have replaced $\widetilde{\sigma}_{jk}^2$ with $\|f\|_\infty$ in the definition of the threshold. But this requires a strong assumption: f is bounded and $\|f\|_\infty$ is known. In our paper, $\mathrm{Var}(\hat{\beta}_{jk})$ is accurately estimated making these conditions unnecessary. Theoretically, we slightly overestimate σ_{jk}^2 to control large deviation terms and this is the reason why we introduce $\widetilde{\sigma}_{jk}^2$. Note that Reynaud-Bouret and Rivoirard (2010) have proposed thresholding rules based on similar heuristic arguments in the Poisson intensity estimation framework. But proofs and computations are more involved for density estimation because sharp upper and lower bounds for $\widetilde{\sigma}_{jk}^2$ are more intricate.

For practical purpose, $\eta_{jk,\gamma}$ (even with $\gamma = 1$) slightly oversmooths the estimate with respect to η_{jk}^{Prac}. From a simulation point of view, the linear term $\frac{2\|\psi_{jk}\|_\infty \log n}{3n}$ in η_{jk}^{Prac} with the precise constant $2/3$ seems to be accurate.

The remaining part of this section is dedicated to a precise choice of γ, first from an oracle point of view, next from a theoretical and practical study.

2.1. Oracle inequalities

Our main result is the following.

Theorem 1. *Let us consider a biorthogonal wavelet basis satisfying the properties described in Appendix A. If $\gamma > c$, then $\tilde{f}_{n,\gamma}$ satisfies the following inequality: for n large enough*

$$\mathbb{E}\left[\|\tilde{f}_{n,\gamma} - f\|_2^2\right] \le C_1 \left[\sum_{(j,k)\in\Gamma_n} \min\left(\beta_{jk}^2, \log n \frac{\sigma_{jk}^2}{n}\right) + \sum_{(j,k)\notin\Gamma_n} \beta_{jk}^2\right] + \frac{C_2 \log n}{n} \tag{2.10}$$

where C_1 is a positive constant depending only on γ, c and the choice of the wavelet basis and where C_2 is also a positive constant depending on γ, c, c', $\|f\|_2$ and the choice of the wavelet basis.

As claimed before, Theorem 1 holds with $c = 1$ and $\gamma > 1$. This result is also true provided $f \in \mathbb{L}_2(\mathbb{R})$. So, assumptions on f are very mild here. This is not the case for most of the results for non-parametric estimation procedures where one assumes that $\|f\|_\infty < \infty$ and that f has a compact support. Furthermore, note that this support and $\|f\|_\infty$ are often known in the literature. On the contrary, in Theorem 1, f and its support can be unbounded. So, we make as few assumptions as possible. This is allowed by considering random thresholding with the data-driven thresholds defined in (2.5).

This result is actually an oracle inequality from the thresholding or the model selection point of view. Indeed, if we consider for each deterministic subset of indices m of Γ_n, the estimator $\hat{f}_m = \sum_{(j,k)\in m} \hat{\beta}_{jk}\tilde{\psi}_{jk}$, we easily see that $\mathbb{E}\left[\|\hat{f}_m - f\|_2^2\right] \asymp R_{\ell_2}(m)$ (see (A.1) in Appendix A for the precise multiplicative constants), with

$$R_{\ell_2}(m) = \sum_{(j,k)\notin m} \beta_{jk}^2 + \sum_{(j,k)\in m} \frac{\sigma_{jk}^2}{n}.$$

Hence the best possible set of indices corresponds to $\overline{m}$ with

$$\overline{m} = \left\{(j,k) \in \Gamma_n \ \text{ such that } \ \beta_{jk}^2 > \frac{\sigma_{jk}^2}{n}\right\}$$

since $\overline{m}$ minimizes $m \mapsto R_{\ell_2}(m)$ and we have

$$R_{\ell_2}(\overline{m}) = \sum_{(j,k)\in\Gamma_n} \min\left(\beta_{jk}^2, \frac{\sigma_{jk}^2}{n}\right) + \sum_{(j,k)\notin\Gamma_n} \beta_{jk}^2.$$

We can associate to $\overline{m}$ the oracle $\hat{f}_{\overline{m}}$, which is not an estimator since it depends on f. Nevertheless, it represents the benchmark in the family of estimators that keep or kill each coefficient $\hat{\beta}_{jk}$. This is exactly the oracle point of view introduced by Donoho and Johnstone (1994) adapted to the density setting. With this approach, we see that Theorem 1 provides the best possible inequality up to a logarithmic term and a residual term. From a thresholding point of view, this logarithmic term is unavoidable as it can be seen when minimax rates are established on the maxisets of thresholding rules derived from such oracle inequalities (See Reynaud-Bouret and Rivoirard (2010) in the Poisson setting for further details). It can also be viewed as the price we pay for not having any information on the coefficients to keep.

With the model selection approach proposed by Birgé and Massart (2007), we can see that Theorem 1 implies

$$\mathbb{E}\left[\|\tilde{f}_{n,\gamma} - f\|_2^2\right] \le C \log(n) \inf_{0<L\le+\infty} \inf_{m\in\mathcal{M}_L} \mathbb{E}\left[\|\hat{f}_m - f\|_2^2\right] + \frac{C_2 \log n}{n},$$

where C is a constant and $\mathcal{M}_L$ represents all the possible sets m in Γ_n such that $\hat{f}_m$ has support in $[-L, L]$. So Theorem 1 consists also in an oracle inequality for estimators assuming that f has a (known) finite support. Finally let us remark that Theorem 1 also implies

$$\mathbb{E}\left[\|\tilde{f}_{n,\gamma} - f\|_2^2\right] \le C \inf_{0<L\le+\infty} \inf_{m\in\mathcal{M}_L} \left\{\sum_{(j,k)\notin m} \beta_{jk}^2 + \frac{|m|\log(n)}{n}\|f\|_\infty\right\} + \frac{C_2 \log n}{n},$$

where $|m|$ is the cardinal of the set m. Of course, this inequality makes sense only if $\|f\|_\infty < \infty$ (see Birgé (2008) for the capital role of $\|f\|_\infty$ when oracle inequalities involve models dimension). This inequality is analogous to the oracle inequality proved by Birgé and Massart (2007) for complex families (such as $\mathcal{M}_L$) in the Gaussian setup. Birgé and Massart also proved that for such families the logarithmic term is unavoidable.

2.2. Calibration issues

We address the problem of choosing conveniently the tuning parameter γ from the theoretical point of view. The aim and the proofs are inspired by Birgé and Massart (2007) who considered penalized estimators and calibrated constants for penalties in a Gaussian framework. In particular, they showed that if the penalty constant is smaller than 1, then the penalized estimator behaves in a quite unsatisfactory way. This study was used in practice to derive adequate data-driven penalties by Lebarbier (2005).

According to Theorem 1, we notice that for any signal, taking $c = 1$ and $c' = 0$, we achieve the oracle performance up to a logarithmic term provided $\gamma > 1$. So, our primary interest is to wonder what happens, from the theoretical point of view, when $\gamma \leq 1$?

To handle this problem, we consider the simplest signal in our setting and we compare the rates of convergence when $\gamma > 1$ and $\gamma < 1$.

Theorem 2. *Let $f = \mathbb{1}_{[0,1]}$ and let us consider $\tilde{f}_{n,\gamma}$ with the Haar basis, $c = 1$ and $c' = 0$.*

- *If $\gamma > 1$ then there exists a constant C depending only on γ such that*

$$\mathbb{E}\left[\|\tilde{f}_{n,\gamma} - f\|_2^2\right] \leq C\frac{\log n}{n}.$$

- *If $\gamma < 1$, then there exists $\delta < 1$ depending only on γ such that*

$$\mathbb{E}\left[\|\tilde{f}_{n,\gamma} - f\|_2^2\right] \geq \frac{1}{n^\delta}(1 + o_n(1)).$$

Theorem 2 establishes that, asymptotically, $\tilde{f}_{n,\gamma}$ with $\gamma < 1$ cannot estimate a very simple signal ($f = \mathbb{1}_{[0,1]}$) at a convenient rate of convergence. This provides a lower bound for the tuning parameter γ: we have to take $\gamma \geq 1$.

We reinforce these results by a simulation study. First we simulate 1000 n-samples of density $f = \mathbb{1}_{[0,1]}$. We estimate f by $\tilde{f}_n^{Prac}$ using the Haar basis, but to see the influence of the parameter γ on the estimation, we replace η_{jk}^{Prac} (see **Step 2** (2.6)) by

$$\eta_{jk} = \sqrt{2\gamma\widehat{\sigma}_{jk}^2\frac{\log n}{n}} + \frac{2\gamma\|\psi_{jk}\|_\infty \log n}{3n}. \tag{2.11}$$

For any γ, we have computed $MISE_n(\gamma)$ i.e. the average over the 1000 simulations of $\|\tilde{f}_n^{Prac} - f\|_2^2$. On the left part of Figure 1 (**U**), $MISE_n(\gamma)\times n$ is plotted as a function of γ for different values of n. Note that when $\gamma > 1$, $MISE_n(\gamma)$ is null meaning that our procedure selects just one wavelet coefficient, the one associated to $\psi_{-1,0} = \mathbb{1}_{[0,1]}$; all others are equal to zero. This fact remains true for a very large range of values of γ. This plateau phenomenon has already been noticed in the Poisson framework (see Reynaud-Bouret and Rivoirard (2010)). However as soon as $\gamma < 1$, $MISE_n(\gamma) \times n$ is positive and increases when γ decreases. It also increases with n tending to prove that $MISE_n(\gamma) >> 1/n$ for $\gamma < 1$. This is in complete adequation with Theorem 2. Remark that, from a theoretical point of view, the proof of part 2 of Theorem 2 holds for any choice of threshold that is asymptotically equivalent to $\sqrt{2\gamma\widehat{\sigma}_{jk}^2\frac{\log n}{n}}$ in the heavy mass zone and

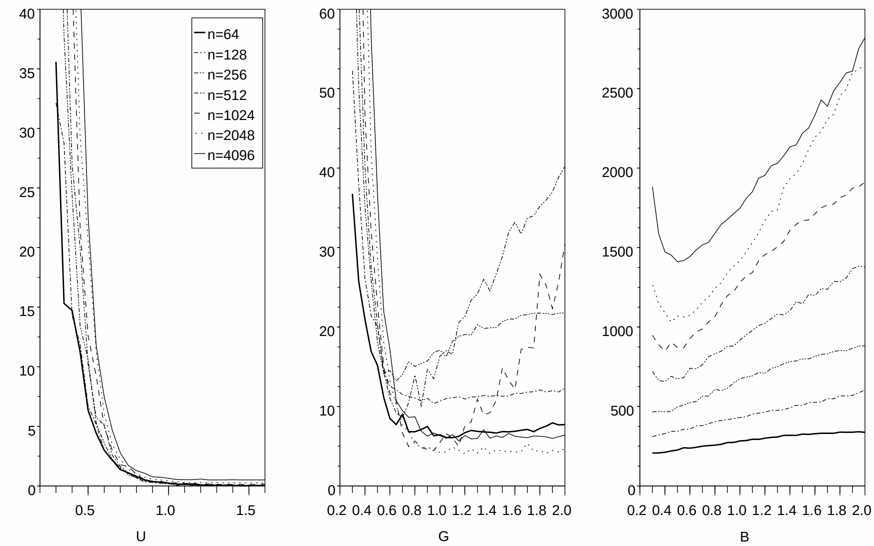

Figure 1: $n \times MISE_n(\gamma)$ for (**U**) $f = \mathbb{1}_{[0,1]}$ (the Haar basis is used) ; (**G**) f is the Gaussian density with mean 0.5 and standard deviation 0.25 (the Spline basis is used) ; (**B**) f is the renormalized Bumps signal (the Spline basis is used)

in particular for the choice (2.11). From a numerical point of view, the left part of Figure 1 (**U**) would have been essentially the same with $\eta_{jk,\gamma}$, i.e. (2.5) instead of (2.11). The reason why we used (2.11) is the practical performance when the function f is more irregular with respect to the chosen basis. Indeed we consider two other density functions f. The first one is the density of a Gaussian variable whose results appear in the middle part of Figure 1 (**G**) and the second one is the renormalized Bumps signal [1] whose results appear in the right part of Figure 1 (**B**). In both cases we computed $\tilde{f}_n^{Prac}$ with the Spline basis: this basis is a particular possible choice of the wavelet basis which leads to smooth estimates. A description is available in Figure 9 of Appendix A. We computed the associate $MISE_n(\gamma)$ over 100 simulations. Note that for the Bumps signal, there is no plateau phenomenon and that the best choice for γ is $\gamma = 0.5$ as soon as the highest level of resolution, $j_0(n)$ is high enough to capture the irregularity of the signal. If n is too small, the best choice is to keep all the coefficients. As already noticed in Reynaud-Bouret and Rivoirard (2010), there exists in fact two behaviors: either the oracle $\hat{f}_{\bar{m}}$ is close to f and the best possible choice is $\gamma \simeq 1$ with a plateau phenomenon, or the oracle $\hat{f}_{\bar{m}}$ is far from f and it is better to take a smaller γ (for instance $\gamma = 0.5$). The Gaussian density (**G**) exhibits both behaviors. For large n ($n \geq 1024$), there is a plateau phenomenon around $\gamma = 1$. But for smaller n, the oracle $\hat{f}_{\bar{m}}$ is not accurate enough and taking $\gamma = 0.5$ is better. Note finally that the choice $\gamma = 1$, leading to our practical method, namely $\tilde{f}_n^{Prac}$, is the more robust with respect to both situations.

[1] The renormalized Bumps signal is a very irregular signal that is classically used in wavelet analysis. It is here renormalized so that the integral equals 1 and it can be defined by $\left(\sum_j g_j \left(1 + \frac{|x-p_j|}{w_j}\right)^{-4}\right) \frac{\mathbf{1}_{[0,1]}}{0.284}$ with

p	=	[	0.1	0.13	0.15	0.23	0.25	0.4	0.44	0.65	0.76	0.78	0.81	]
g	=	[	4	5	3	4	5	4.2	2.1	4.3	3.1	5.1	4.2	]
w	=	[	0.005	0.005	0.006	0.01	0.01	0.03	0.01	0.01	0.005	0.008	0.005	]

3. The curse of support from a minimax point of view

The goal of this section is to derive the minimax rates on the whole class of Besov spaces. The subsequent results will constitute generalizations of the results derived in Juditsky and Lambert-Lacroix (2004) who pointed out minimax rates for density estimation on the class of Hölder spaces. For this purpose, we consider the theoretical procedure $\tilde{f}_{n,\gamma}$ defined with the choice $c' = -c$ (see **Step 0**) where the real number c is chosen later. In some situations, it will be necessary to strengthen our assumptions. More precisely, sometimes, we assume that f is bounded. So, for any $R > 0$, we consider the following set of functions:

$$\mathcal{L}_{2,\infty}(R) = \{f \text{ is a density such that } \|f\|_2 \leq R \text{ and } \|f\|_\infty \leq R\}.$$

The Besov balls we consider are classical (see Appendix A for a definition with respect to the biorthogonal wavelet basis) and denoted $\mathcal{B}^\alpha_{p,q}(R)$. Let us just point out that no restriction is made on the support of f when f belongs to $\mathcal{B}^\alpha_{p,q}(R)$: this support is potentially the whole real line. Now, let us state the upper bound of the $\mathbb{L}_2$-risk of $\tilde{f}_{n,\gamma}$.

Theorem 3. *Let $R, R' > 0$, $1 \leq p, q \leq \infty$ and $\alpha \in \mathbb{R}$ such that $\max\left(0, \frac{1}{p} - \frac{1}{2}\right) < \alpha < r + 1$, where $r > 0$ denotes the wavelet smoothness parameter introduced in Appendix A. Let $c \geq 1$ such that*

$$\alpha\left(1 - \frac{1}{c(1+2\alpha)}\right) \geq \frac{1}{p} - \frac{1}{2} \tag{3.1}$$

and $\gamma > c$. Then, there exists a constant C depending on R', γ, c, α, p and on the choice of the biorthogonal wavelet basis such that for n large enough,

- *if $p \leq 2$,*
$$\sup_{f \in \mathcal{B}^\alpha_{p,q}(R) \cap \mathcal{L}_{2,\infty}(R')} \mathbb{E}\left[\|\tilde{f}_{n,\gamma} - f\|_2^2\right] \leq C R^{\frac{2}{2\alpha+1}} \left(\frac{\log n}{n}\right)^{\frac{2\alpha}{2\alpha+1}}, \tag{3.2}$$
- *if $p > 2$,*
$$\sup_{f \in \mathcal{B}^\alpha_{p,q}(R) \cap \mathbb{L}_2(R')} \mathbb{E}\left[\|\tilde{f}_{n,\gamma} - f\|_2^2\right] \leq C R^{\frac{1}{\alpha+1-\frac{1}{p}}} \left(\frac{\log n}{n}\right)^{\frac{\alpha}{\alpha+1-\frac{1}{p}}}, \tag{3.3}$$

  where here $\mathbb{L}_2(R')$ denote the set of densities such that $\|f\|_2 \leq R'$.

First, let us briefly comment assumptions of these results. When $p > 2$, (3.1) is satisfied and the result is true for any $c \geq 1$ and $0 < \alpha < r+1$. Furthermore, we do not need to restrict ourselves to the set of bounded functions. When $p \leq 2$, the result is true as soon as c is large enough to satisfy (3.1) and we establish (3.2) only for bounded functions. Actually, this assumption is in some sense unavoidable as proved in Section 6.4 of Birgé (2008).

Furthermore, note that if we additionally assume that f is bounded with a bounded support (say $[0, 1]$) then $\mathbb{E}\left[\|\tilde{f}_{n,\gamma} - f\|_2^2\right]$ is always upper bounded by a constant times $(\log n/n)^{\frac{2\alpha}{2\alpha+1}}$ whatever p is, since, in this case when $p > 2$, the assumption $f \in \mathcal{B}^\alpha_{p,\infty}(R)$ implies $f \in \mathcal{B}^\alpha_{2,\infty}(\tilde{R})$ for $\tilde{R}$ large enough.

Now, combining upper bounds (3.2) and (3.3), under assumptions of Theorem 3, we point out the following rate for our procedure when f is bounded but without any assumption on the support: for n large enough,

$$\sup_{f \in \mathcal{B}^\alpha_{p,q}(R) \cap \mathcal{L}_{2,\infty}(R')} \mathbb{E}\left[\|\tilde{f}_{n,\gamma} - f\|_2^2\right] \leq C R^{\frac{1}{\alpha+\frac{1}{2}+\left(\frac{1}{2}-\frac{1}{p}\right)_+}} \left(\frac{\log n}{n}\right)^{\frac{\alpha}{\alpha+\frac{1}{2}+\left(\frac{1}{2}-\frac{1}{p}\right)_+}}.$$

The following result derives lower bounds of the minimax risk showing that this rate is the optimal rate up to a logarithmic term. So, the next result establishes the optimality properties of $\tilde{f}_{n,\gamma}$ under the minimax approach.

Theorem 4. *Let $R, R' > 0$, $1 \leq p, q \leq \infty$ and $\alpha \in \mathbb{R}$ such that* $\max\left(0, \frac{1}{p} - \frac{1}{2}\right) < \alpha < r + 1$. *Then, there exists a positive constant $\tilde{C}$ depending on R', α, p and q such that*

$$\liminf_{n \to +\infty} n^{\frac{\alpha}{\alpha + \frac{1}{2} + \left(\frac{1}{2} - \frac{1}{p}\right)_+}} \inf_{\hat{f}} \sup_{f \in \mathcal{B}^{\alpha}_{p,q}(R) \cap \mathcal{L}_{2,\infty}(R')} \mathbb{E}\left[\|\hat{f} - f\|_2^2\right] \geq \tilde{C} R^{\frac{1}{\alpha + \frac{1}{2} + \left(\frac{1}{2} - \frac{1}{p}\right)_+}},$$

where the infimum is taken over all the possible density estimators $\hat{f}$.
Furthermore, let c, $p^* \geq 1$ and $\alpha^* > 0$ such that

$$\alpha^* \left(1 - \frac{1}{c(1 + 2\alpha^*)}\right) \geq \frac{1}{p^*} - \frac{1}{2}. \tag{3.4}$$

Then our procedure, $\tilde{f}_{n,\gamma}$, constructed with this precise choice of c and $\gamma > c$, is adaptive minimax up to a logarithmic term on

$$\left\{\mathcal{B}^{\alpha}_{p,q}(R) \cap \mathcal{L}_{2,\infty}(R') : \quad \alpha^* \leq \alpha < r + 1,\ p^* \leq p \leq +\infty,\ 1 \leq q \leq \infty\right\}.$$

When $p \leq 2$, the lower bound for the minimax risk corresponds to the classical minimax rate for estimating a compactly supported density (see Donoho *et al.* (1996)). Furthermore, the procedure $\tilde{f}_{n,\gamma}$ achieves this minimax rate up to a logarithmic term. When $p > 2$, the risk deteriorates, if no assumption on the support is made, whereas it remains the same when we add the bounded support assumption. Note that when $p = \infty$, the exponent becomes $\alpha/(1+\alpha)$: it was also derived in Juditsky and Lambert-Lacroix (2004) for estimation on balls of $\mathcal{B}^{\alpha}_{\infty,\infty}$. We also mention that when $p \geq 2$, convenient non-adaptive linear estimates achieve the optimal rate without logarithmic term. It is a simple consequence of technical arguments used for proving Theorem 2 of Juditsky and Lambert-Lacroix (2004).

To summarize, we gather in Table 1 the lower bounds for the minimax rates obtained for each situation. These bounds are adaptively achieved by our estimator with respect to p, α and the compactness of the support, up to a logarithmic term.

	$1 \leq p \leq 2$	$2 \leq p \leq \infty$
compact support	$n^{-\frac{2\alpha}{2\alpha+1}}$	$n^{-\frac{2\alpha}{2\alpha+1}}$
non compact support	$n^{-\frac{2\alpha}{2\alpha+1}}$	$n^{-\frac{\alpha}{\alpha+1-\frac{1}{p}}}$

Table 1: Minimax rates on $\mathcal{B}^{\alpha}_{p,q} \cap \mathcal{L}_{2,\infty}$(up to a logarithmic term) with $1 \leq p, q \leq \infty$, $\alpha > \max\left(0, \frac{1}{p} - \frac{1}{2}\right)$ under the $\|\cdot\|_2^2$-loss.

Our results show the role played by the support of the functions to be estimated on minimax rates. As already observed, when $p \leq 2$, the support has no influence since the rate exponent remains unchanged whatever the size of the support (finite or not). Roughly speaking, it means that it is not harder to estimate bounded non-compactly supported functions than bounded compactly supported functions from the minimax point of view. It is not the case when $p > 2$. Actually, we note an elbow phenomenon at $p = 2$ and the rate deteriorates when p increases: this illustrates the curse of support from a minimax point of view. Let us give an interpretation of this observation. Johnstone (1994) showed that when $p < 2$, Besov spaces $\mathcal{B}^{\alpha}_{p,q}$ model sparse signals where at each level, a very few number of the wavelet coefficients are non-negligible. But these coefficients can be very large. When $p > 2$, $\mathcal{B}^{\alpha}_{p,q}$-spaces typically model dense signals where the wavelet coefficients are not large but most of them can be non-negligible. This explains why the size of the support plays a role on minimax rates when $p > 2$: when the support is larger, the number of wavelet coefficients to be estimated increases dramatically.

Since arguments for proving Theorems 3 and 4 are similar to the arguments used in Reynaud-Bouret and Rivoirard (2010), proofs are omitted. We just mention that Theorem 3 is derived from the oracle inequality established in Theorem 1.

Finally, a natural interesting extension of this work could be to investigate rates for $\mathbb{L}_{p'}$-loss functions, $1 \leq p' < \infty$. Note that the case $p' = \infty$ is very different in nature (see Giné and Nickl (2009) and Giné and Nickl (2010)).

4. The curse of support from a practical point of view

Now let us turn to a practical point of view. Is there a curse of support too? First we provide a simulation study illustrating the distortion of the most classic support dependent estimators when the support or the tail is increasing. Next we provide an application of our method to famous real data sets, namely the Suicide data and the Old Faithful geyser data.

4.1. Simulations

We compare our method to representative methods of each main trend in density estimation, namely kernel, binning plus thresholding and model selection. The considered methods are the following. The first one is the kernel method, denoted **K**, consisting in a basic cross-validation choice of a global bandwidth with a Gaussian kernel. The second method requires a complex preprocessing of the data based on binning. Observations $X_1, \ldots, X_n$ are first rescaled and centered by an affine transformation denoted T such that $T(X_1), \ldots, T(X_n)$ lie in $[0, 1]$. We denote f_T the density of the data induced by the transformation T. We divide the interval $[0, 1]$ into 2^{b_n} small intervals of size 2^{-b_n}, where b_n is an integer, and count the number of observations in each interval. We apply the root transform due to Brown *et al.* (2010) and the universal hard individual thresholding rule on the coefficients computed with the DWT Coiflet-basis filter. We finally apply the unroot transform to obtain an estimate of f_T and the final estimate of the density is obtained by applying T^{-1} combined with a spline interpolation. This method is denoted **RU**. The last method is also support dependent. After rescaling as previously the data, we estimate f_T by the algorithm of Willett and Nowak (2007). It consists in a complex selection of a grid and of polynomials on that grid that minimizes a penalized log-likelihood criterion. The final estimate of the density is obtained by applying T^{-1}. This method is denoted **WN**.
Our practical method is implemented in the Haar basis (method **H**) and in the Spline basis (method **S**) (see Figure 9 in Appendix A for a complete description of this basis). Moreover we have also implemented the choice $\gamma = 0.5$ of (2.11) in the Spline basis (see Section 2). We denote this method **S***.
The thresholding rule proposed in Juditsky and Lambert-Lacroix (2004) has also been considered. For their prescribed practical choice of the tuning parameters and the Spline basis, the numerical performances are similar to those of method **S**. Since thresholding is not performed for the coarsest level, the approximation term of the reconstruction is based on many non zero negligible coefficients for heavy-tailed signals: this leads to obvious numerical difficulties without significant impact on the risk. So, numerical results of the thresholding rule proposed in Juditsky and Lambert-Lacroix (2004) are not given in the sequel.

We generate n-samples of two kinds of densities f, with $n = 1024$. Both signals are supported by the whole real line. We compute for each estimator $\hat{f}$ the ISE, i.e. $\int_{\mathbb{R}}(f - \hat{f})^2$ which is approximated by a trapezoidal method on a finite interval, adequately chosen so that the remaining term is negligible with respect to the ISE.

The first signal, g_d, consists in a mixture of two standard Gaussian densities:

$$g_d = \frac{1}{2}\,\mathcal{N}(0,1) + \frac{1}{2}\,\mathcal{N}(d,1),$$

where $\mathcal{N}(\mu,\sigma)$ represents the density of a Gaussian variable with mean μ and standard deviation σ. The parameter d varies in $\{10,30,50,70\}$ so that we can see the curse of support on the quality of estimation.

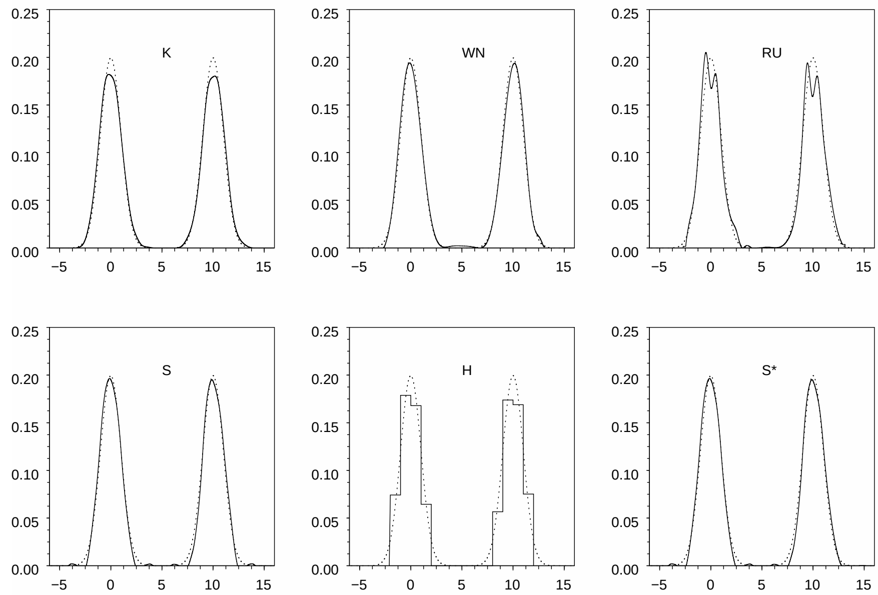

Figure 2: Reconstruction of g_d (true: dotted line, estimate: solid line) for the 6 different methods for $d = 10$

Figure 2 shows the reconstructions for $d = 10$ and Figure 3 for $d = 70$. In the sequel, the method **RU** is implemented with $b_n = 5$, which is the best choice for the reconstruction with $d = 10$. All the methods give satisfying results for $d = 10$. When d is large, the rescaling and binning preprocessing leads to a poor regression signal which makes the regression thresholding rules non convenient, as illustrated by the method **RU** with $d = 70$. Reconstructions for **K, WN, S** and **S*** seem satisfying but a study of the ISE of each method (see Figure 4) reveals that both support dependent methods (**RU** and **WN**) have a risk that increases with d. On the contrary, methods **K** and **S** are the best ones and more interestingly their performance is remarkably stable (the boxsize is quite small) and the result does not vary with d. This robustness is also true for **H** and **S***. **S*** is a bit undersmoothing: this was already noticed in Figure 1 (**G**) and this explains the variability of its ISE. Finally note that, for large d, **H** is even better than **RU** despite the inappropriate choice of the Haar basis.

The other signal, h_k, is both heavy-tailed and irregular. It consists in a mixture of 4 Gaussian densities and one Student density:

$$h_k = 0.45\,T(k) + 0.15\,\mathcal{N}(-1,0.05) + 0.1\,\mathcal{N}(-0.7,0.005) + 0.25\,\mathcal{N}(1,0.025) + 0.15\,\mathcal{N}(2,0.05),$$

where $T(k)$ denotes the density of a Student variable with k degrees of freedom. The parameter k varies in $\{2,4,8,16\}$. The smaller k, the heavier the tail is and this without changing the shape of the main part that has to be estimated. Figure 5 shows the reconstruction for $k = 2$. Clearly **RU** does not detect the local spikes at all. Indeed the maximal observation may be equal to 1000 and the binning effect is disastrous.

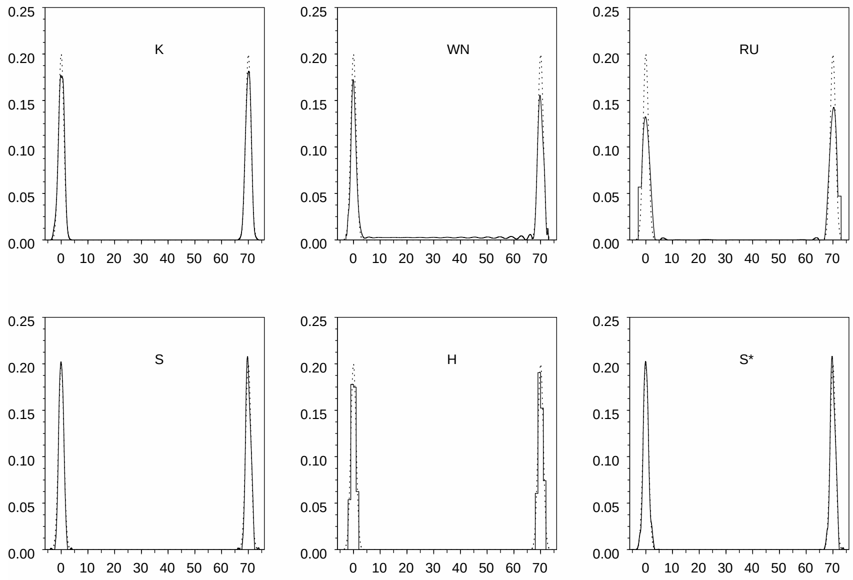

Figure 3: Reconstruction of g_d (true: dotted line, estimate: solid line) for the 6 different methods for $d = 70$

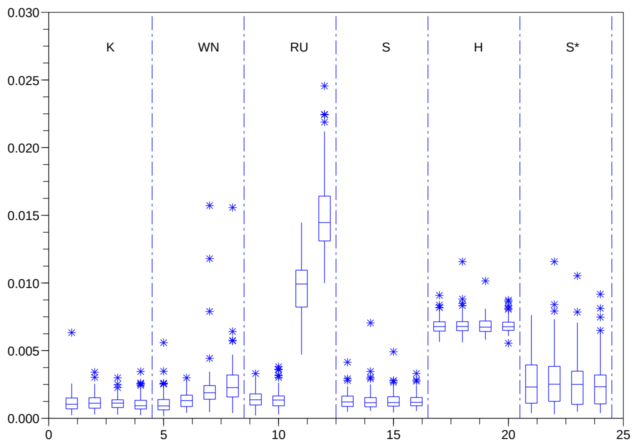

Figure 4: Boxplots of the ISE for g_d over 100 simulations for the 6 methods and the 4 different values of d. A column, delimited by dashed lines, corresponds to one method (respectively **K, WN, RU, S, H, S***). Inside this column, from left to right, one can find for the same method the boxplots of the ISE for respectively $d = 10, 30, 50$ and 70.

The kernel method **K** clearly suffers from a lack of spatial adaptivity, as expected. The four remaining methods seem satisfying. In particular for this very irregular signal it is not clear that the Haar basis is a bad choice. Note however that to represent reconstructions, we have focused on the area where the spikes are located. In particular we emphasize that Figure 5 does not show that the support dependent method **WN** is non zero on a very large interval, which tends to deteriorate its ISE. Indeed, Figure 6 shows that the ISE of the support dependent methods (**RU, WN**) increases when the tail becomes heavier, whereas the other methods have remarkable stable ISE. Methods **S** and **H** are more robust and better than **WN** for $k = 2$. The ISE may be improved for this irregular signal by taking $\gamma = 0.5$ (see method **S***) as already noticed in Section 2 for irregular signals.

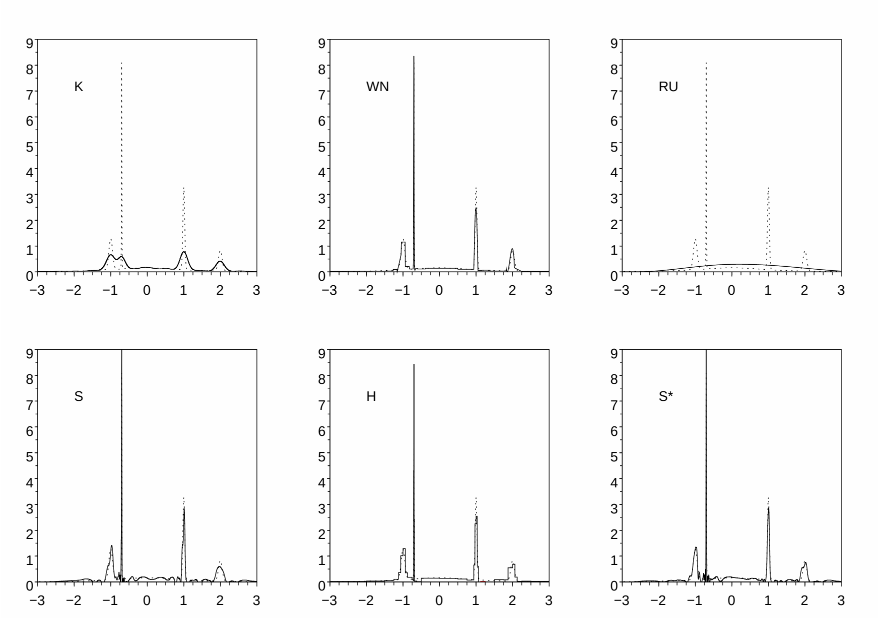

Figure 5: Reconstruction of h_k (true: dotted line, estimate: solid line) for the 6 different methods for $k = 2$

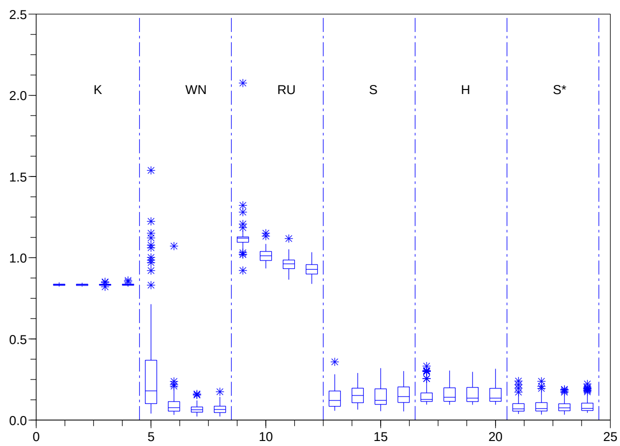

Figure 6: Boxplots of the ISE for h_k over 100 simulations for the 6 methods and the 4 different values of k. A column, delimited by dashed lines, corresponds to one method (respectively **K, WN, RU, S, H, S***). Inside this column, from left to right, one can find for the same method the boxplots of the ISE for respectively $k = 2, 4, 8$ and 16.

4.2. On real data

To illustrate and evaluate our procedure on real data, we consider two real data sets named, respectively in our study, “Old Faithful geyser” and “Suicide”. The “Old Faithful geyser” data are the duration, in minutes, of 107 eruptions of Old Faithful geyser located in Yellowstone National Park, USA; they are taken from Weisberg (1980). The “Suicide” data set is related to the study of suicide risks. Indeed, each of the 86 observations corresponds to the number of days a patient, considered as control in the study, undergoes psychiatric treatment. The data are available in Copas and Fryer (1980). In both cases, we consider that we have a sample of n real observations $X_1, \ldots, X_n$ and we want to estimate the underlying density f. We mention that in the first situation, all the observations are continuous whereas, in the second one, the observations are discrete. These data are well known and have been widely studied elsewhere. This allows to compare our procedure with other methods.

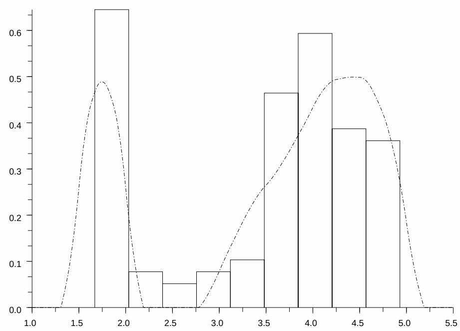

Figure 7: Histogram (solid line) and reconstruction via $\tilde{f}_n^{Prac}$ (dashed line) for the ”Old Faithful geyser” data set

To estimate the function f, we apply $\tilde{f}_n^{Prac}$, with the Spline basis (see Figure 9 in Appendix A) and $j_0 = 7$. We plot, on the same graph the resulting estimate and the histogram of the data. Figures 7 and 8 represent, respectively, the results for the “Old Faithful geyser” set and for the “Suicide” one. Note that concerning the ”Suicide” data set, there exists a problem of ”scale”: if we look at the associated histogram, the scale of the data seems to be approximately equal to 250, and not 1. So we divide the data by 250 before performing estimation.
Respectively two or three peaks are detected providing multimodal reconstructions. So, in comparison with the ones performed in Silverman (1986) and Sain and Scott (1996), our estimate detects significant events and not artefacts. More interestingly, with few observations, both estimates equal zero on an interval located between the last two peaks. Even if it is hard to build a good estimate of the true density due to the small number of the data, the advantage of having this “hole” is to provide a good separation between both modes. Note that a Gaussian kernel estimate, which is never null, provides sharp mode localization only when the bandwidth is small enough but in this case, the kernel estimate becomes noisy for heavy tailed data (see Silverman (1986) p.18). On the contrary, when $\tilde{f}_n^{Prac}$ is null, this does not mean that the true density is null but only negligible. If the practitioner keeps this fact in mind, then $\tilde{f}_n^{Prac}$ provides a good interpretation of real data even for small sample size.

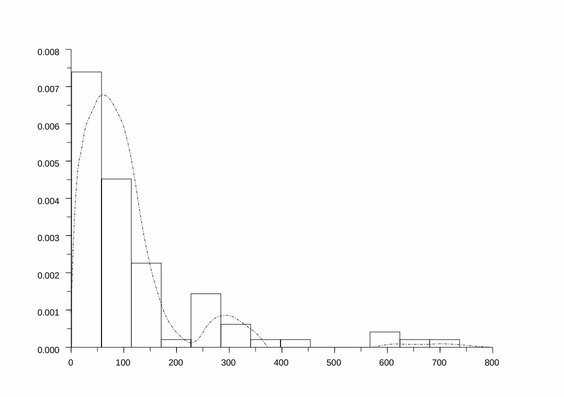

Figure 8: Histogram (solid line) and reconstruction via $\tilde{f}_n^{Prac}$ (dashed line) for the "Suicide" data set

A. Analytical tools

All along this paper, we have considered a particular class of wavelet bases that are described now. We set

$$\phi = \mathbb{1}_{[0,1]}.$$

For any $r > 0$, we can claim that there exist three functions ψ, $\tilde{\phi}$ and $\tilde{\psi}$ with the following properties:

1. $\tilde{\phi}$ and $\tilde{\psi}$ are compactly supported,
2. $\tilde{\phi}$ and $\tilde{\psi}$ belong to C^{r+1}, where C^{r+1} denotes the Hölder space of order $r + 1$,
3. ψ is compactly supported and is a piecewise constant function,
4. ψ is orthogonal to polynomials of degree no larger than r,
5. $\{(\phi_k, \psi_{jk})_{j\geq 0,k\in\mathbb{Z}}, (\tilde{\phi}_k, \tilde{\psi}_{jk})_{j\geq 0,k\in\mathbb{Z}}\}$ is a biorthogonal family: for any $j, j' \geq 0$, for any k, k',

$$\int_{\mathbb{R}} \psi_{jk}(x)\tilde{\phi}_{k'}(x)dx = \int_{\mathbb{R}} \phi_k(x)\tilde{\psi}_{j'k'}(x)dx = 0,$$

$$\int_{\mathbb{R}} \phi_k(x)\tilde{\phi}_{k'}(x)dx = 1_{k=k'}, \quad \int_{\mathbb{R}} \psi_{jk}(x)\tilde{\psi}_{j'k'}(x)dx = 1_{j=j',k=k'},$$

where for any $x \in \mathbb{R}$,

$$\phi_k(x) = \phi(x-k), \quad \psi_{jk}(x) = 2^{\frac{j}{2}}\psi(2^j x - k)$$

and

$$\tilde{\phi}_k(x) = \tilde{\phi}(x-k), \quad \tilde{\psi}_{jk}(x) = 2^{\frac{j}{2}}\tilde{\psi}(2^j x - k).$$

This implies the following wavelet decomposition of $f \in \mathbb{L}_2(\mathbb{R})$:

$$f = \sum_{k\in\mathbb{Z}} \alpha_k \tilde{\phi}_k + \sum_{j\geq 0}\sum_{k\in\mathbb{Z}} \beta_{jk}\tilde{\psi}_{jk},$$

where for any $j \geq 0$ and any $k \in \mathbb{Z}$,

$$\alpha_k = \int_{\mathbb{R}} f(x)\phi_k(x)dx, \quad \beta_{jk} = \int_{\mathbb{R}} f(x)\psi_{jk}(x)dx.$$

Such biorthogonal wavelet bases have been built by Cohen *et al.* (1992) as a special case of spline systems (see also the elegant equivalent construction of Donoho (1994) from boxcar functions). The Haar basis can

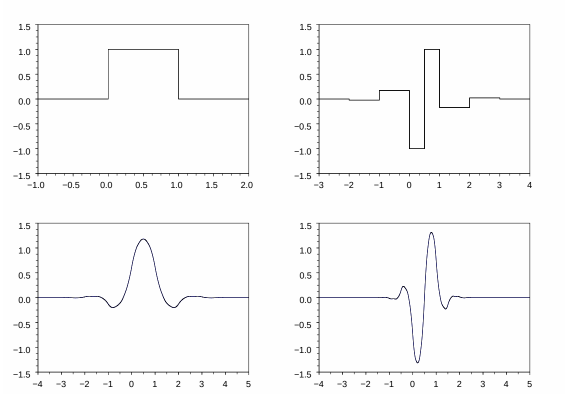

Figure 9: Biorthogonal wavelet basis with $r = 0.272$ that is used in the Simulation study. First line, ϕ (left) and ψ (right), second line $\tilde{\phi}$ (left) and $\tilde{\psi}$ (right).

be viewed as a particular biorthogonal wavelet basis, by setting $\tilde{\phi} = \phi$ and $\tilde{\psi} = \psi = \mathbb{1}_{[0,\frac{1}{2})} - \mathbb{1}_{[\frac{1}{2},1]}$, with $r = 0$ (even if Property 2 is not satisfied with such a choice). The Haar basis is an orthonormal basis, which is not true for general biorthogonal wavelet bases. However, we have the frame property: if we denote

$$\Phi = \{\phi, \psi, \tilde{\phi}, \tilde{\psi}\}$$

there exist two constants $c_1(\Phi)$ and $c_2(\Phi)$ only depending on Φ such that

$$c_1(\Phi)\left(\sum_{k\in\mathbb{Z}} \alpha_k^2 + \sum_{j\geq 0}\sum_{k\in\mathbb{Z}} \beta_{jk}^2\right) \leq \|f\|_2^2 \leq c_2(\Phi)\left(\sum_{k\in\mathbb{Z}} \alpha_k^2 + \sum_{j\geq 0}\sum_{k\in\mathbb{Z}} \beta_{jk}^2\right). \tag{A.1}$$

For instance, when the Haar basis is considered, $c_1(\Phi) = c_2(\Phi) = 1$.

We emphasize the important feature of such bases: the functions ψ_{jk} are piecewise constant functions. For instance, Figure 9 shows an example which is the one that has been implemented for numerical studies. This allows to compute easily wavelet coefficients without using the discrete wavelet transform. Furthermore, there exists a constant $\mu_\psi > 0$ such that

$$\inf_{x\in[0,1]} |\phi(x)| \geq 1, \qquad \inf_{x\in \mathrm{Supp}(\psi)} |\psi(x)| \geq \mu_\psi,$$

where $\mathrm{Supp}(\psi) = \{x \in \mathbb{R} : \quad \psi(x) \neq 0\}$.

This technical feature will be used through the proofs of our results. To shorten mathematical expressions, we have previously set for any $k \in \mathbb{Z}$, $\tilde{\psi}_{-1k} = \tilde{\phi}_k$, $\psi_{-1k} = \phi_k$ and $\beta_{-1k} = \alpha_k$.

Now, let us give some properties of Besov spaces. Besov spaces, denoted $\mathcal{B}^\alpha_{p,q}$, are classically defined by using modulus of continuity (see DeVore and Lorentz (1993) and Härdle *et al.* (1998)). We just recall here the sequential characterization of Besov spaces by using the biorthogonal wavelet basis (for further details, see Delyon and Juditsky (1997)).

Let $1 \leq p, q \leq \infty$ and $0 < \alpha < r + 1$, the $\mathcal{B}^\alpha_{p,q}$-norm of f is equivalent to the norm

$$\|f\|_{\alpha,p,q} = \begin{cases} \|(\alpha_k)_k\|_{\ell_p} + \left[\sum_{j\geq 0} 2^{jq(\alpha+\frac{1}{2}-\frac{1}{p})}\|(\beta_{j,k})_k\|^q_{\ell_p}\right]^{1/q} & \text{if } q < \infty, \\ \|(\alpha_k)_k\|_{\ell_p} + \sup_{j\geq 0} 2^{j(\alpha+\frac{1}{2}-\frac{1}{p})}\|(\beta_{j,k})_k\|_{\ell_p} & \text{if } q = \infty. \end{cases}$$

We use this norm to define Besov balls with radius R

$$\mathcal{B}^\alpha_{p,q}(R) = \{f \in \mathbb{L}_2(\mathbb{R}) : \quad \|f\|_{\alpha,p,q} \leq R\}.$$

For any $R > 0$, if $0 < \alpha' \leq \alpha < r+1$, $1 \leq p \leq p' \leq \infty$ and $1 \leq q \leq q' \leq \infty$, we obviously have

$$\mathcal{B}^{\alpha}_{p,q}(R) \subset \mathcal{B}^{\alpha}_{p,q'}(R), \quad \mathcal{B}^{\alpha}_{p,q}(R) \subset \mathcal{B}^{\alpha'}_{p,q}(R).$$

Moreover

$$\mathcal{B}^{\alpha}_{p,q}(R) \subset \mathcal{B}^{\alpha'}_{p',q}(R) \text{ if } \alpha - \frac{1}{p} \geq \alpha' - \frac{1}{p'}.$$

The class of Besov spaces provides a useful tool to classify wavelet decomposed signals with respect to their regularity and sparsity properties (see Johnstone (1994)). Roughly speaking, regularity increases when α increases whereas sparsity increases when p decreases.

B. Proofs

B.1. Proof of Theorem 1

Because of the frame property of the biorthogonal wavelet basis, it is easy to see that

$$c_1(\Phi)\|\tilde{\beta} - \beta\|^2_{\ell_2} \leq \|\tilde{f}_{n,\gamma} - f\|^2_2 \leq c_2(\Phi)\|\tilde{\beta} - \beta\|^2_{\ell_2}, \tag{B.1}$$

where $\tilde{\beta}$ denotes the sequence of thresholded coefficients $(\tilde{\beta}_{jk}\mathbb{1}_{(j,k)\in\Gamma_n})_{(j,k)\in\Lambda}$ and β denotes the true coefficients $(\beta_{jk})_{(j,k)\in\Lambda}$. Consequently, it is sufficient to restrict ourselves to the study of the $\|\tilde{\beta} - \beta\|^2_{\ell_2}$.

Consequently the proof of Theorem 1 relies on the following result (see Theorem 7 of Section 4.1 in Reynaud-Bouret and Rivoirard (2010)).

Theorem 5. *Let Λ be a set of indices. To estimate a countable family $\beta = (\beta_\lambda)_{\lambda\in\Lambda}$ such that $\|\beta\|_{\ell_2} < \infty$, we assume that a family of coefficient estimators $(\hat{\beta}_\lambda)_{\lambda\in\Gamma}$, where Γ is a known deterministic subset of Λ, and a family of possibly random thresholds $(\eta_\lambda)_{\lambda\in\Gamma}$ are available and we consider the thresholding rule $\tilde{\beta} = (\hat{\beta}_\lambda\mathbb{1}_{|\hat{\beta}_\lambda|\geq\eta_\lambda}\mathbb{1}_{\lambda\in\Gamma})_{\lambda\in\Lambda}$. Let $\varepsilon > 0$ be fixed. Assume that there exist a deterministic family $(F_\lambda)_{\lambda\in\Gamma}$ and three constants $\kappa \in [0, 1[$, $\omega \in [0, 1]$ and $\mu > 0$ (that may depend on ε but not on λ) with the following properties.*

(A1) For all $\lambda \in \Gamma$,

$$\mathbb{P}(|\hat{\beta}_\lambda - \beta_\lambda| > \kappa\eta_\lambda) \leq \omega.$$

(A2) There exist $1 < p, q < \infty$ with $\frac{1}{p} + \frac{1}{q} = 1$ and a constant $R > 0$ such that for all $\lambda \in \Gamma$,

$$\left(\mathbb{E}(|\hat{\beta}_\lambda - \beta_\lambda|^{2p})\right)^{\frac{1}{p}} \leq R\max(F_\lambda, F_\lambda^{\frac{1}{p}}\varepsilon^{\frac{1}{q}}).$$

(A3) There exists a constant θ such that for all $\lambda \in \Gamma$ satisfying $F_\lambda < \theta\varepsilon$

$$\mathbb{P}(|\hat{\beta}_\lambda - \beta_\lambda| > \kappa\eta_\lambda\ , |\hat{\beta}_\lambda| > \eta_\lambda) \leq F_\lambda\mu.$$

Then the estimator $\tilde{\beta}$ satisfies

$$\frac{1-\kappa^2}{1+\kappa^2}\mathbb{E}\|\tilde{\beta} - \beta\|^2_{\ell_2} \leq \mathbb{E}\inf_{m\subset\Gamma}\left\{\frac{1+\kappa^2}{1-\kappa^2}\sum_{\lambda\notin m}\beta^2_\lambda + \frac{1-\kappa^2}{\kappa^2}\sum_{\lambda\in m}(\hat{\beta}_\lambda - \beta_\lambda)^2 + \sum_{\lambda\in m}\eta^2_\lambda\right\} + LD\sum_{\lambda\in\Gamma}F_\lambda$$

with

$$LD = \frac{R}{\kappa^2}\left(\left(1 + \theta^{-1/q}\right)\omega^{1/q} + (1 + \theta^{1/q})\varepsilon^{1/q}\mu^{1/q}\right).$$

To prove Theorem 1, we use Theorem 5 with $\lambda = (j,k)$, $\hat{\beta}_\lambda = \hat{\beta}_{jk}$ defined in (2.4), $\eta_{jk} = \eta_{jk,\gamma}$ defined in (2.5) and

$$\Gamma = \Gamma_n = \{(j,k) \in \Lambda : \ -1 \le j \le j_0\} \text{ with } 2^{j_0} \le n^c (\log n)^{c'} < 2^{j_0+1}.$$

We set

$$F_{jk} = \int_{\mathrm{Supp}(\psi_{jk})} f(x) dx.$$

Hence we have:

$$\sum_{(j,k)\in\Gamma_n} F_{jk} = \sum_{-1\le j\le j_0} \sum_k \int_{x\in \mathrm{Supp}(\psi_{jk})} f(x)dx \le \int f(x)dx \sup_{x\in\mathbb{R}} \left[\sum_{-1\le j\le j_0} \sum_k \mathbb{1}_{x\in\mathrm{Supp}(\psi_{jk})} \right] \le (j_0+2)m_\psi, \quad \text{(B.2)}$$

where m_ψ is a finite constant depending only on the compactly supported function ψ. Finally, $\sum_{(j,k)\in\Gamma_n} F_{jk}$ is bounded by $\log(n)$ up to a constant that only depends on c, c' and the function ψ. Now, we give a fundamental lemma to derive Assumption (A1) of Theorem 5.

Lemma 1. *For any $\gamma > 1$ and any $\varepsilon' > 0$ there exists a constant M depending on ε' and γ such that*

$$\mathbb{P}\left(\sigma^2_{jk} \ge (1+\varepsilon')\widetilde{\sigma}^2_{jk}\right) \le M n^{-\gamma}.$$

Proof. We have:

$$\begin{aligned}
\widetilde{\sigma}^2_{jk} &= \frac{1}{2n(n-1)} \sum_{i\ne l} (\psi_{jk}(X_i) - \psi_{jk}(X_l))^2 \\
&= \frac{1}{2n}\sum_{i=1}^n (\psi_{jk}(X_i) - \beta_{jk})^2 + \frac{1}{2n}\sum_{l=1}^n (\psi_{jk}(X_l) - \beta_{jk})^2 \\
&\quad - \frac{2}{n(n-1)} \sum_{i=2}^n \sum_{l=1}^{i-1} (\psi_{jk}(X_i) - \beta_{jk})(\psi_{jk}(X_l) - \beta_{jk}) \\
&= s_n - \frac{2}{n(n-1)} u_n
\end{aligned} \quad \text{(B.3)}$$

with

$$s_n = \frac{1}{n}\sum_{i=1}^n (\psi_{jk}(X_i) - \beta_{jk})^2 \quad \text{and} \quad u_n = \sum_{i=2}^n \sum_{l=1}^{i-1} (\psi_{jk}(X_i) - \beta_{jk})(\psi_{jk}(X_l) - \beta_{jk}).$$

Using the Bernstein inequality (see section 2.2.3 in Massart (2007)) applied to the variables Y_i with

$$Y_i = \frac{\sigma^2_{jk} - (\psi_{jk}(X_i) - \beta_{jk})^2}{n} \le \frac{\sigma^2_{jk}}{n},$$

one obtains for any $u > 0$,

$$\mathbb{P}\left(\sigma^2_{jk} \ge s_n + \sqrt{2v_{jk}u} + \frac{\sigma^2_{jk}u}{3n}\right) \le e^{-u}$$

with

$$v_{jk} = \frac{1}{n}\mathbb{E}\left[\left(\sigma^2_{jk} - (\psi_{jk}(X_i) - \beta_{jk})^2\right)^2\right].$$

We have

$$\begin{aligned}
v_{jk} &= \frac{1}{n}\left(\sigma^4_{jk} + \mathbb{E}\left[(\psi_{jk}(X_i) - \beta_{jk})^4\right] - 2\sigma^2_{jk}\mathbb{E}\left[(\psi_{jk}(X_i) - \beta_{jk})^2\right]\right) \\
&= \frac{1}{n}\left(\mathbb{E}\left[(\psi_{jk}(X_i) - \beta_{jk})^4\right] - \sigma^4_{jk}\right) \\
&\le \frac{\sigma^2_{jk}}{n}\left(\|\psi_{jk}\|_\infty + |\beta_{jk}|\right)^2 \\
&\le \frac{4\sigma^2_{jk}}{n}\|\psi_{jk}\|^2_\infty.
\end{aligned}$$

Finally

$$\mathbb{P}\left(\sigma_{jk}^2 \geq s_n + 2\|\psi_{jk}\|_\infty \sigma_{jk}\sqrt{\frac{2u}{n}} + \frac{\sigma_{jk}^2 u}{3n}\right) \leq e^{-u}. \tag{B.4}$$

Now, we deal with the degenerate U-statistics u_n. We use Theorem 3.1 of Houdré and Reynaud-Bouret (2003) combined with the appropriate choice of constants derived by Klein and Rio (2005): for any $u > 0$ and any $\tau > 0$,

$$\mathbb{P}\left(u_n \geq (1+\tau)C\sqrt{2u} + 2Du + \frac{1+\tau}{3}Fu + \left(\sqrt{2}(3+\tau^{-1}) + \frac{2}{3}\right)Bu^{3/2} + \frac{3+\tau^{-1}}{3}Au^2\right) \leq 3e^{-u}. \tag{B.5}$$

Note that similar results with unknown constants have been derived in the seminal work by Giné *et al.* (2000). Here we use a sharper bound.

Now we need to define and control the 5 quantities A, B, C, D and F. For this purpose, let us set for any x and y,

$$g_{jk}(x,y) = (\psi_{jk}(x) - \beta_{jk})(\psi_{jk}(y) - \beta_{jk}).$$

We have:

$$A = \|g_{jk}\|_\infty \leq 4\|\psi_{jk}\|_\infty^2.$$

Furthermore,

$$C^2 = \sum_{i=2}^{n}\sum_{l=1}^{i-1}\mathbb{E}(g_{jk}^2(X_i, X_l)) = \frac{n(n-1)}{2}\sigma_{jk}^4.$$

The next term is

$$\begin{aligned}
D &= \sup_{\mathbb{E}\sum a_i^2(X_i)\leq 1,\ \mathbb{E}\sum b_l^2(X_l)\leq 1} \mathbb{E}\left(\sum_{i=2}^{n}\sum_{l=1}^{i-1} g_{jk}(X_i, X_l)a_i(X_i)b_l(X_l)\right)\\
&= \sup_{\mathbb{E}\sum a_i^2(X_i)\leq 1,\ \mathbb{E}\sum b_l^2(X_l)\leq 1} \sum_{i=2}^{n}\sum_{l=1}^{i-1}\mathbb{E}\Big((\psi_{jk}(X_i) - \beta_{jk})a_i(X_i)\Big)\mathbb{E}\Big((\psi_{jk}(X_l) - \beta_{jk})b_l(X_l)\Big)\\
&\leq \sup_{\mathbb{E}\sum a_i^2(X_i)\leq 1,\ \mathbb{E}\sum b_l^2(X_l)\leq 1} \sum_{i=2}^{n}\sum_{l=1}^{i-1}\sqrt{\sigma_{jk}^2\mathbb{E}(a_i^2(X_i))}\sqrt{\sigma_{jk}^2\mathbb{E}(b_l^2(X_l))}.
\end{aligned}$$

So, we have

$$\begin{aligned}
D &\leq \sigma_{jk}^2 \sup_{\mathbb{E}\sum a_i^2(X_i)\leq 1,\ \mathbb{E}\sum b_l^2(X_l)\leq 1} \sum_{i=2}^{n}\sqrt{\mathbb{E}(a_i^2(X_i))}\sqrt{\sum_{l=1}^{i-1}\mathbb{E}(b_l^2(X_l))}\sqrt{i-1}\\
&\leq \sigma_{jk}^2 \sup_{\mathbb{E}\sum a_i^2(X_i)\leq 1} \sqrt{\sum_{i=2}^{n}\mathbb{E}(a_i^2(X_i))}\sqrt{\sum_{i=2}^{n}(i-1)}\\
&\leq \sigma_{jk}^2\sqrt{\frac{n(n-1)}{2}}.
\end{aligned}$$

Still using Theorem 3.1 of Houdré and Reynaud-Bouret (2003), we have:

$$\begin{aligned}
B^2 &= \sup_t \sum_{l=1}^{n-1}\mathbb{E}((\psi_{jk}(t) - \beta_{jk})^2(\psi_{jk}(X_l) - \beta_{jk})^2)\\
&\leq 4(n-1)\|\psi_{jk}\|_\infty^2\sigma_{jk}^2\\
&\leq 4(n-1)\|\psi_{jk}\|_\infty^4
\end{aligned}$$

Finally

$$
\begin{aligned}
F &= \mathbb{E}\left(\sup_{i,t}\left|\sum_{l=1}^{i-1}(\psi_{jk}(t)-\beta_{jk})(\psi_{jk}(X_l)-\beta_{jk})\right|\right)\\
&\leq 2\|\psi_{jk}\|_\infty \mathbb{E}\left(\sup_{i}\left|\sum_{l=1}^{i-1}(\psi_{jk}(X_l)-\beta_{jk})\right|\right).
\end{aligned}
$$

To control this term, we set

$$
Z_i=\sum_{l=1}^{i-1}(\psi_{jk}(X_l)-\beta_{jk}).
$$

Using Doob's inequality on the martingale $(Z_i)_i$, we obtain

$$
\mathbb{E}(\sup_i |Z_i|)\leq \sqrt{\mathbb{E}(\sup_i Z_i^2)}\leq 2\sup_i \sqrt{\mathbb{E}(Z_i^2)}=2\sigma_{jk}\sqrt{n-1}.
$$

Hence

$$
F\leq 4\|\psi_{jk}\|_\infty\sigma_{jk}\sqrt{n-1}.
$$

Now, for any $u>0$, let us set

$$
S(u)=2\|\psi_{jk}\|_\infty\sigma_{jk}\sqrt{2\frac{u}{n}}+\frac{\sigma_{jk}^2 u}{3n}
$$

and

$$
U(u)=(1+\tau)C\sqrt{2u}+2Du+\frac{1+\tau}{3}Fu+\left(\sqrt{2}(3+\tau^{-1})+\frac{2}{3}\right)Bu^{3/2}+\frac{3+\tau^{-1}}{3}Au^2.
$$

Inequalities (B.4) and (B.5) give

$$
\begin{aligned}
\mathbb{P}\left(\sigma_{jk}^2\geq\widehat{\sigma}_{jk}^2+S(u)+\frac{2}{n(n-1)}U(u)\right) &= \mathbb{P}\left(\sigma_{jk}^2\geq s_n+S(u)+\frac{2}{n(n-1)}(U(u)-u_n)\right)\\
&\leq \mathbb{P}\left(\sigma_{jk}^2\geq s_n+S(u)\right)+\mathbb{P}(u_n\geq U(u))\\
&\leq 4e^{-u}.
\end{aligned}
$$

Let us take $u=\gamma\log n$ and $\tau=1$. Then, there exist some constants a and b depending on γ such that

$$
S(u)+\frac{2}{n(n-1)}U(u)\leq 2\sigma_{jk}\|\psi_{jk}\|_\infty\sqrt{2\gamma\frac{\log n}{n}}+a\sigma_{jk}^2\frac{\log n}{n}+b\|\psi_{jk}\|_\infty^2\left(\frac{\log n}{n}\right)^{3/2}.
$$

So,

$$
\mathbb{P}\left(\sigma_{jk}^2\geq\widehat{\sigma}_{jk}^2+2\sigma_{jk}\|\psi_{jk}\|_\infty\sqrt{2\gamma\frac{\log n}{n}}+a\sigma_{jk}^2\frac{\log n}{n}+b\|\psi_{jk}\|_\infty^2\left(\frac{\log n}{n}\right)^{3/2}\right)\leq 4n^{-\gamma}
$$

and

$$
\mathbb{P}\left(\sigma_{jk}^2\left(1-a\frac{\log n}{n}\right)-2\sigma_{jk}\|\psi_{jk}\|_\infty\sqrt{2\gamma\frac{\log n}{n}}-\widehat{\sigma}_{jk}^2-b\|\psi_{jk}\|_\infty^2\left(\frac{\log n}{n}\right)^{3/2}\geq 0\right)\leq 4n^{-\gamma}.
$$

Now, we set

$$
\theta_1=\left(1-a\frac{\log n}{n}\right),\quad \theta_2=\|\psi_{jk}\|_\infty\sqrt{2\gamma\frac{\log n}{n}}
$$

and

$$
\theta_3=\widehat{\sigma}_{jk}^2+b\|\psi_{jk}\|_\infty^2\left(\frac{\log n}{n}\right)^{3/2}
$$

with $\theta_1, \theta_2, \theta_3 > 0$ for n large enough depending only on γ. We study the polynomial

$$p(\sigma) = \theta_1\sigma^2 - 2\theta_2\sigma - \theta_3.$$

Then, since $\sigma \geq 0$, $p(\sigma) \geq 0$ means that

$$\sigma \geq \frac{1}{\theta_1}\left(\theta_2 + \sqrt{\theta_2^2 + \theta_1\theta_3}\right),$$

which is equivalent to

$$\sigma^2 \geq \frac{1}{\theta_1^2}\left(2\theta_2^2 + \theta_1\theta_3 + 2\theta_2\sqrt{\theta_2^2 + \theta_1\theta_3}\right).$$

Hence

$$\mathbb{P}\left(\sigma_{jk}^2 \geq \frac{1}{\theta_1^2}\left(2\theta_2^2 + \theta_1\theta_3 + 2\theta_2\sqrt{\theta_2^2 + \theta_1\theta_3}\right)\right) \leq 4n^{-\gamma}.$$

So,

$$\mathbb{P}\left(\sigma_{jk}^2 \geq \frac{\theta_3}{\theta_1} + \frac{2\theta_2\sqrt{\theta_3}}{\theta_1\sqrt{\theta_1}} + \frac{4\theta_2^2}{\theta_1^2}\right) \leq 4n^{-\gamma}.$$

So, there exist absolute constants δ, η, and τ' depending only on γ so that for n large enough,

$$\mathbb{P}\left(\sigma_{jk}^2 \geq \widehat{\sigma}_{jk}^2\left(1+\delta\frac{\log n}{n}\right) + \left(1+\eta\frac{\log n}{n}\right)2\|\psi_{jk}\|_\infty\sqrt{2\gamma\widehat{\sigma}_{jk}^2\frac{\log n}{n}} + 8\gamma\|\psi_{jk}\|_\infty^2\frac{\log n}{n}\left(1+\tau'\left(\frac{\log n}{n}\right)^{1/4}\right)\right) \leq 4n^{-\gamma}.$$

Hence, with

$$\widetilde{\sigma}_{jk}^2 = \widehat{\sigma}_{jk}^2 + 2\|\psi_{jk}\|_\infty\sqrt{2\gamma\widehat{\sigma}_{jk}^2\frac{\log n}{n}} + 8\gamma\|\psi_{jk}\|_\infty^2\frac{\log n}{n},$$

for all $\varepsilon' > 0$ there exists M such that

$$\mathbb{P}(\sigma_{jk}^2 \geq (1+\varepsilon')\widetilde{\sigma}_{jk}^2) \leq Mn^{-\gamma}.$$

■

Let $\kappa < 1$. Applying the previous lemma gives

$$\begin{aligned}
\mathbb{P}(|\hat{\beta}_{jk} - \beta_{jk}| > \kappa\eta_{jk,\gamma}) &\leq \mathbb{P}\left(|\hat{\beta}_{jk} - \beta_{jk}| \geq \sqrt{2\kappa^2\gamma\widetilde{\sigma}_{jk}^2\frac{\log n}{n}} + \frac{2\kappa\gamma\log n\|\psi_{jk}\|_\infty}{3n}\right)\\
&\leq \mathbb{P}\left(|\hat{\beta}_{jk} - \beta_{jk}| \geq \sqrt{2\kappa^2\gamma\widetilde{\sigma}_{jk}^2\frac{\log n}{n}} + \frac{2\kappa\gamma\log n\|\psi_{jk}\|_\infty}{3n},\ \sigma_{jk}^2 \geq (1+\varepsilon')\widetilde{\sigma}_{jk}^2\right)\\
&\quad + \mathbb{P}\left(|\hat{\beta}_{jk} - \beta_{jk}| \geq \sqrt{2\kappa^2\gamma\widetilde{\sigma}_{jk}^2\frac{\log n}{n}} + \frac{2\kappa\gamma\log n\|\psi_{jk}\|_\infty}{3n},\ \sigma_{jk}^2 < (1+\varepsilon')\widetilde{\sigma}_{jk}^2\right)\\
&\leq \mathbb{P}\left(\sigma_{jk}^2 \geq (1+\varepsilon')\widetilde{\sigma}_{jk}^2\right)\\
&\quad + \mathbb{P}\left(|\hat{\beta}_{jk} - \beta_{jk}| \geq \sqrt{2\kappa^2\gamma(1+\varepsilon')^{-1}\sigma_{jk}^2\frac{\log n}{n}} + \frac{2\kappa\gamma\log n\|\psi_{jk}\|_\infty}{3n}\right).
\end{aligned}$$

Using again the Bernstein inequality, we have for any $u > 0$,

$$\mathbb{P}\left(|\hat{\beta}_{jk} - \beta_{jk}| \geq \sqrt{\frac{2u\sigma_{jk}^2}{n}} + \frac{2u\|\psi_{jk}\|_\infty}{3n}\right) \leq 2e^{-u}.$$

So, with $\varepsilon' = 1 - \kappa$, there exists a constant M_κ depending only on κ and γ such that

$$\mathbb{P}(|\hat{\beta}_{jk} - \beta_{jk}| > \kappa\eta_{jk,\gamma}) \leq M_\kappa n^{-\gamma\kappa^2/(2-\kappa)}.$$

So, for any value of $\kappa \in (0,1[$, Assumption (A1) is true with $\eta_{jk} = \eta_{jk,\gamma}$ if we take $\omega = M_\kappa n^{-\gamma\kappa^2/(2-\kappa)}$.
Now, to prove (A2), we use the Rosenthal inequality. There exists a constant $C(p)$ only depending on p such that

$$\begin{aligned}
\mathbb{E}(|\hat{\beta}_{jk} - \beta_{jk}|^{2p}) &= \frac{1}{n^{2p}}\mathbb{E}\left[\left|\sum_{i=1}^{n}\left(\psi_{jk}(X_i) - \mathbb{E}(\psi_{jk}(X_i))\right)\right|^{2p}\right] \\
&\leq \frac{C(p)}{n^{2p}}\left(\sum_{i=1}^{n}\mathbb{E}\left[\left|\psi_{jk}(X_i) - \mathbb{E}(\psi_{jk}(X_i))\right|^{2p}\right] + \left(\sum_{i=1}^{n}\mathrm{Var}(\psi_{jk}(X_i))\right)^{p}\right)
\end{aligned}$$

$$\begin{aligned}
\mathbb{E}(|\hat{\beta}_{jk} - \beta_{jk}|^{2p}) &\leq \frac{C(p)}{n^{2p}}\left(\sum_{i=1}^{n}\left(2\|\psi_{jk}\|_\infty\right)^{2p-2}\mathrm{Var}(\psi_{jk}(X_i)) + \left(\sum_{i=1}^{n}\mathrm{Var}(\psi_{jk}(X_i))\right)^{p}\right) \\
&\leq \frac{C(p)}{n^{2p}}\left(\left(2\|\psi_{jk}\|_\infty\right)^{2p-2} n\sigma_{jk}^2 + n^p\sigma_{jk}^{2p}\right) \\
&\leq \frac{C(p)}{n^{2p}}\left(\left(2\|\psi_{jk}\|_\infty\right)^{2p} nF_{jk} + n^p\|\psi_{jk}\|_\infty^{2p}F_{jk}^p\right).
\end{aligned}$$

Finally,

$$\begin{aligned}
\left(\mathbb{E}(|\hat{\beta}_{jk} - \beta_{jk}|^{2p})\right)^{\frac{1}{p}} &\leq \frac{4C(p)^{\frac{1}{p}}\|\psi_{jk}\|_\infty^2}{n}\left(n^{1-p}F_{jk} + F_{jk}^p\right)^{\frac{1}{p}} \\
&\leq \frac{4C(p)^{\frac{1}{p}}2^{j_0}\max(\|\phi\|_\infty^2;\|\psi\|_\infty^2)}{n}\left(n^{-\frac{1}{q}}F_{jk}^{\frac{1}{p}} + F_{jk}\right).
\end{aligned}$$

So, Assumption (A2) is satisfied with $\varepsilon = \frac{1}{n}$ and

$$R = \frac{8C(p)^{\frac{1}{p}}2^{j_0}\max(\|\phi\|_\infty^2;\|\psi\|_\infty^2)}{n}.$$

Finally, to prove Assumption (A3), we use the following lemma.

Lemma 2. *We set*

$$N_{jk} = \sum_{i=1}^{n}\mathbb{1}_{\{X_i \in \mathrm{Supp}(\psi_{jk})\}} \quad \textit{and} \quad C' = \frac{14\gamma}{3} \geq \frac{14}{3}.$$

There exists an absolute constant $0 < \theta' < 1$ *such that if* $nF_{jk} \leq \theta' C' \log n$ *and* $(1-\theta')\log n \geq \frac{3}{7}$ *then,*

$$\mathbb{P}(N_{jk} - nF_{jk} \geq (1-\theta')C'\log n) \leq F_{jk}n^{-\gamma}.$$

Proof. One takes $\theta' \in [0,1]$ such that

$$\frac{(1-\theta')^2}{(2\theta'+1)} \geq \frac{4}{7}.$$

We use the Bernstein inequality that yields

$$\mathbb{P}(N_{jk} - nF_{jk} \geq (1-\theta')C'\log n) \leq \exp\left(-\frac{((1-\theta')C'\log n)^2}{2(nF_{jk} + (1-\theta')C'\log n/3)}\right) \leq n^{-\frac{3C'(1-\theta')^2}{2(2\theta'+1)}}.$$

If $nF_{jk} \geq n^{-\gamma-1}$, since $\frac{3C'(1-\theta')^2}{2(2\theta'+1)} \geq 2\gamma+2$, the result is true. If $nF_{jk} \leq n^{-\gamma-1}$, using properties of Binomial random variables (see page 482 of Shorack and Wellner (1986)), for $n \geq 2$,

$$\begin{aligned}
\mathbb{P}(N_{jk} - nF_{jk} \geq (1-\theta')C'\log n) \leq \mathbb{P}(N_{jk} > (1-\theta')C'\log n) &\leq \mathbb{P}(N_{jk} \geq 2) \\
&\leq \frac{(1-F_{jk})C_n^2F_{jk}^2(1-F_{jk})^{n-2}}{1-3^{-1}(n+1)F_{jk}} \\
&\leq \frac{n^2F_{jk}^2}{2(1-2^{-1}nF_{jk})} \\
&\leq (nF_{jk})^2
\end{aligned}$$

and the result is true. ∎

Now, observe that if $|\hat{\beta}_{jk}| > \eta_{jk,\gamma}$ then

$$N_{jk} \geq C' \log n.$$

Indeed, $|\hat{\beta}_{jk}| > \eta_{jk,\gamma}$ implies

$$\frac{C' \log n}{n} \|\psi_{jk}\|_\infty \leq |\hat{\beta}_{jk}| \leq \frac{\|\psi_{jk}\|_\infty N_{jk}}{n}.$$

So, if n satisfies $(1-\theta')\log n \geq \frac{3}{7}$, we set $\theta = \theta' C' \log(n)$ and $\mu = n^{-\gamma}$. In this case, Assumption (A3) is fulfilled since if $nF_{jk} \leq \theta' C' \log n$

$$\mathbb{P}(|\hat{\beta}_{jk} - \beta_{jk}| > \kappa \eta_{jk,\gamma}, |\hat{\beta}_{jk}| > \eta_{jk,\gamma}) \leq \mathbb{P}(N_{jk} - nF_{jk} \geq (1-\theta')C' \log n) \leq F_{jk} n^{-\gamma}.$$

Finally, if n satisfies $(1-\theta')\log n \geq \frac{3}{7}$, we can apply Theorem 5 and we have:

$$\frac{1-\kappa^2}{1+\kappa^2}\mathbb{E}\|\tilde{\beta}-\beta\|_{\ell_2}^2 \leq \inf_{m \subset \Gamma_n}\left\{\frac{1+\kappa^2}{1-\kappa^2}\sum_{(j,k)\notin m}\beta_{jk}^2 + \frac{1-\kappa^2}{\kappa^2}\sum_{(j,k)\in m}\mathbb{E}(\hat{\beta}_{jk}-\beta_{jk})^2 + \sum_{(j,k)\in m}\mathbb{E}(\eta_{jk,\gamma}^2)\right\} + LD\sum_{(j,k)\in\Gamma_n} F_{jk}. \tag{B.6}$$

Furthermore, there exists a constant K_1 depending on p, γ, κ, c, c' and on ψ such that

$$LD \sum_{(j,k)\in\Gamma_n} F_{jk} \leq K_1 (\log(n))^{c'+1} n^{c - \frac{\kappa^2\gamma}{q(2-\kappa)} - 1}. \tag{B.7}$$

Since $\gamma > c$, one takes $0 < \kappa < 1$ and $q > 1$ such that $c < \frac{\kappa^2\gamma}{q(2-\kappa)}$ and as required by Theorem 1, the last term satisfies

$$LD \sum_{(j,k)\in\Gamma_n} F_{jk} \leq \frac{K_2}{n},$$

where K_2 is a constant. Now we can derive the oracle inequality. Before evaluating the first term of (B.6), let us state the following lemma.

Lemma 3. *We set for any $(j,k) \in \Lambda$*

$$D_{jk} = \int \psi_{jk}^2(x) f(x) dx,$$

$$S_\psi = \max\{\sup_{x\in \mathrm{Supp}(\phi)} |\phi(x)|, \sup_{x\in\mathrm{Supp}(\psi)} |\psi(x)|\}$$

and

$$I_\psi = \min\{\inf_{x\in \mathrm{Supp}(\phi)} |\phi(x)|, \inf_{x\in\mathrm{Supp}(\psi)} |\psi(x)|\}.$$

Using Appendix A, we define $\Theta_\psi = \frac{S_\psi^2}{I_\psi^2}$. For all $(j,k) \in \Lambda$, we have the following result.

- *If $F_{jk} \leq \Theta_\psi \frac{\log(n)}{n}$, then $\beta_{jk}^2 \leq \Theta_\psi^2 D_{jk} \frac{\log(n)}{n}$.*

- *If $F_{jk} > \Theta_\psi \frac{\log(n)}{n}$, then $\|\psi_{jk}\|_\infty \frac{\log(n)}{n} \leq \sqrt{\frac{D_{jk}\log(n)}{n}}$.*

Proof. We assume that $j \geq 0$ (arguments are similar for $j = -1$).
If $F_{jk} \leq \Theta_\psi \frac{\log(n)}{n}$, we have

$$|\beta_{jk}| \leq S_\psi 2^{\frac{j}{2}} F_{jk} \leq S_\psi 2^{\frac{j}{2}} \sqrt{F_{jk}} \sqrt{\Theta_\psi} \sqrt{\frac{\log(n)}{n}} \leq S_\psi I_\psi^{-1} \sqrt{\Theta_\psi} \sqrt{\frac{D_{jk}\log(n)}{n}} \leq \Theta_\psi \sqrt{\frac{D_{jk}\log(n)}{n}},$$

since $D_{jk} \geq I_\psi^2 2^j F_{jk}$. For the second point, observe that

$$\sqrt{\frac{D_{jk}\log(n)}{n}} \geq 2^{\frac{j}{2}} I_\psi \sqrt{\Theta_\psi}\frac{\log(n)}{n} = 2^{\frac{j}{2}} S_\psi \frac{\log(n)}{n} \geq \|\psi_{jk}\|_\infty \frac{\log(n)}{n}.$$

■

Now, for any $\delta > 0$,

$$\mathbb{E}(\eta_{jk,\gamma}^2) \leq (1+\delta)\frac{2\gamma\log n}{n}\mathbb{E}(\widetilde{\sigma}_{jk}^2) + (1+\delta^{-1})\left(\frac{2\gamma\log n}{3n}\right)^2 \|\psi_{jk}\|_\infty^2.$$

Moreover,

$$\frac{\mathbb{E}(\widetilde{\sigma}_{jk}^2)}{n} \leq (1+\delta)\frac{D_{jk}}{n} + (1+\delta^{-1})8\gamma\log n\frac{\|\psi_{jk}\|_\infty^2}{n^2}.$$

So,

$$\mathbb{E}(\eta_{jk,\gamma}^2) \leq (1+\delta)^2 2\gamma\log n\frac{D_{jk}}{n} + \Delta(\delta)\left(\frac{\gamma\log n}{n}\right)^2 \|\psi_{jk}\|_\infty^2, \tag{B.8}$$

with $\Delta(\delta)$ a constant depending only on δ. Now, we apply (B.6) with

$$m = \left\{(j,k) \in \Gamma_n : \quad \beta_{jk}^2 > \Theta_\psi^2 \frac{D_{jk}}{n}\log n\right\},$$

so using Lemma 3, we can claim that for any $(j,k) \in m$, $F_{jk} > \Theta_\psi \frac{\log(n)}{n}$. Finally, since $\Theta_\psi \geq 1$,

$$\begin{aligned}
\mathbb{E}\|\tilde{\beta}-\beta\|_{\ell_2}^2 &\leq K_3\left(\sum_{(j,k)\in\Gamma_n}\beta_{jk}^2 \mathbb{1}_{\{\beta_{jk}^2\leq\Theta_\psi^2\frac{D_{jk}}{n}\log n\}} + \sum_{(j,k)\notin\Gamma_n}\beta_{jk}^2\right)\\
&\quad + K_3\sum_{(j,k)\in\Gamma_n}\left[\frac{\log n}{n}D_{jk} + \left(\frac{\log n}{n}\right)^2 \|\psi_{jk}\|_\infty^2\right]\mathbb{1}_{\left\{\beta_{jk}^2>\Theta_\psi^2\frac{D_{jk}}{n}\log n,\ F_{jk}>\Theta_\psi\frac{\log(n)}{n}\right\}} + \frac{K_4}{n}\\
&\leq K_3\left[\sum_{(j,k)\in\Gamma_n}\left(\beta_{jk}^2\mathbb{1}_{\left\{\beta_{jk}^2\leq\Theta_\psi^2\log n\frac{D_{jk}}{n}\right\}} + 2\log n\frac{D_{jk}}{n}\mathbb{1}_{\left\{\beta_{jk}^2>\Theta_\psi^2\log n\frac{D_{jk}}{n}\right\}}\right) + \sum_{(j,k)\notin\Gamma_n}\beta_{jk}^2\right] + \frac{K_4}{n}\\
&\leq 2K_3\left[\sum_{(j,k)\in\Gamma_n}\min\left(\beta_{jk}^2, \Theta_\psi^2\log n\frac{D_{jk}}{n}\right) + \sum_{(j,k)\notin\Gamma_n}\beta_{jk}^2\right] + \frac{K_4}{n},
\end{aligned}$$

where the constant K_3 depends on γ and c and K_4 depends on γ, c, c' and on ψ. Finally, since

$$D_{jk} = \sigma_{jk}^2 + \beta_{jk}^2,$$

$$\begin{aligned}
\mathbb{E}\|\tilde{\beta}-\beta\|_{\ell_2}^2 &\leq 2K_3\left[\sum_{(j,k)\in\Gamma_n}\min\left(\beta_{jk}^2 + \frac{\Theta_\psi^2\log n}{n}\beta_{jk}^2, \Theta_\psi^2\log n\frac{\sigma_{jk}^2}{n} + \frac{\Theta_\psi^2\log n}{n}\beta_{jk}^2\right) + \sum_{(j,k)\notin\Gamma_n}\beta_{jk}^2\right] + \frac{K_4}{n}\\
&\leq 2K_3\left[\sum_{(j,k)\in\Gamma_n}\min\left(\beta_{jk}^2, \Theta_\psi^2\log n\frac{\sigma_{jk}^2}{n}\right) + \sum_{(j,k)\in\Gamma_n}\frac{\Theta_\psi^2\log n}{n}\beta_{jk}^2 + \sum_{(j,k)\notin\Gamma_n}\beta_{jk}^2\right] + \frac{K_4}{n}\\
&\leq 2K_3\Theta_\psi^2\left[\sum_{(j,k)\in\Gamma_n}\min\left(\beta_{jk}^2, \log n\frac{\sigma_{jk}^2}{n}\right) + \sum_{(j,k)\notin\Gamma_n}\beta_{jk}^2\right] + 2K_3\Theta_\psi^2\|\beta\|_{\ell_2}\frac{\log n}{n} + \frac{K_4}{n}.
\end{aligned}$$

Theorem 1 is proved by using properties of the biorthogonal wavelet basis.

B.2. Proof of Theorem 2

The first part is a direct application of Theorem 1. Now let us turn to the second part. We recall that we consider $f = \mathbb{1}_{[0,1]}$, the Haar basis and for $j \geq 0$ and $k \in \mathbb{Z}$, we have:

$$\widetilde{\sigma}_{jk}^2 = \widehat{\sigma}_{jk}^2 + 2\|\psi_{jk}\|_\infty \sqrt{2\gamma\widehat{\sigma}_{jk}^2 \frac{\log n}{n}} + 8\gamma\|\psi_{jk}\|_\infty^2 \frac{\log n}{n}.$$

So, for any $0 < \varepsilon < \frac{1-\gamma}{2} < \frac{1}{2}$,

$$\widetilde{\sigma}_{jk}^2 \leq (1+\varepsilon)\widehat{\sigma}_{jk}^2 + 2\gamma\|\psi_{jk}\|_\infty^2 \frac{\log n}{n}\left(\varepsilon^{-1} + 4\right).$$

Now,

$$\begin{aligned}
\eta_{jk,\gamma} &= \sqrt{2\gamma\widetilde{\sigma}_{jk}^2 \frac{\log n}{n}} + \frac{2\|\psi_{jk}\|_\infty \gamma \log n}{3n} \\
&\leq \sqrt{2\gamma\frac{\log n}{n}\left((1+\varepsilon)\widehat{\sigma}_{jk}^2 + 2\gamma\|\psi_{jk}\|_\infty^2 \frac{\log n}{n}(\varepsilon^{-1}+4)\right)} + \frac{2\|\psi_{jk}\|_\infty \gamma \log n}{3n} \\
&\leq \sqrt{2\gamma(1+\varepsilon)\widehat{\sigma}_{jk}^2 \frac{\log n}{n}} + \frac{2\|\psi_{jk}\|_\infty \gamma \log n}{n}\left(\frac{1}{3} + \sqrt{4+\varepsilon^{-1}}\right).
\end{aligned}$$

Furthermore, using (B.3)

$$\widehat{\sigma}_{jk}^2 = s_n - \frac{2}{n(n-1)}u_n,$$

and

$$\eta_{jk,\gamma} \leq \sqrt{2\gamma(1+\varepsilon)\frac{\log n}{n}s_n} + \sqrt{2\gamma(1+\varepsilon)\frac{\log n}{n} \times \frac{2}{n(n-1)}|u_n|} + \frac{2\|\psi_{jk}\|_\infty \gamma \log n}{n}\left(\frac{1}{3} + \sqrt{4+\varepsilon^{-1}}\right).$$

Using (B.5), with probability larger than $1 - 6n^{-2}$,

$$|u_n| \leq U(2\log n),$$

and, since $f = \mathbb{1}_{[0,1]}$, we have $\sigma_{jk}^2 \leq 1$ and

$$\frac{2}{n(n-1)}U(2\log n) \quad \leq \quad C_1\frac{\log n}{n} + C_2\|\psi_{jk}\|_\infty^2 \left(\frac{\log n}{n}\right)^{\frac{3}{2}},$$

where C_1 and C_2 are universal constants. Finally, with probability larger than $1 - 6n^{-2}$,

$$\sqrt{2\gamma(1+\varepsilon)\frac{\log n}{n} \times \frac{2}{n(n-1)}|u_n|} \leq \sqrt{2\gamma(1+\varepsilon)C_1}\frac{\log n}{n} + \sqrt{2\gamma(1+\varepsilon)C_2}\|\psi_{jk}\|_\infty \left(\frac{\log n}{n}\right)^{\frac{5}{4}}.$$

So, since $\gamma < 1$, there exists $w(\varepsilon)$, only depending on ε such that with probability larger than $1 - 6n^{-2}$,

$$\eta_{jk,\gamma} \quad \leq \quad \sqrt{2\gamma(1+\varepsilon)\frac{\log n}{n}s_n} + w(\varepsilon)\|\psi_{jk}\|_\infty \frac{\log n}{n}.$$

Since $\|\psi_{jk}\|_\infty = 2^{j/2}$, we set

$$\widetilde{\eta_{jk,\gamma}} = \sqrt{2\gamma(1+\varepsilon)s_n\frac{\log n}{n}} + w(\varepsilon)\frac{2^{\frac{j}{2}}\log n}{n}$$

and $\eta_{jk,\gamma} \leq \widetilde{\eta_{jk,\gamma}}$ with probability larger than $1-6n^{-2}$. Then, since $f = \mathbb{1}_{[0,1]}$, $\beta_{jk} = 0$ for $j \geq 0$ and

$$\begin{aligned} s_n &= \frac{1}{n}\sum_{i=1}^n \left(\psi_{jk}(X_i) - \beta_{jk}\right)^2 \\ &= \frac{2^j}{n}\sum_{i=1}^n \left(\mathbb{1}_{X_i\in[k2^{-j},(k+0.5)2^{-j}[} - \mathbb{1}_{X_i\in[(k+0.5)2^{-j},(k+1)2^{-j}[}\right)^2 \\ &= \frac{2^j}{n}\left(N_{jk}^+ + N_{jk}^-\right), \end{aligned}$$

with

$$N_{jk}^+ = \sum_{i=1}^n \mathbb{1}_{X_i\in[k2^{-j},(k+0.5)2^{-j}[}, \quad N_{jk}^- = \sum_{i=1}^n \mathbb{1}_{X_i\in[(k+0.5)2^{-j},(k+1)2^{-j}[}.$$

We consider j such that

$$\frac{n}{(\log n)^\alpha} \leq 2^j < \frac{2n}{(\log n)^\alpha}, \quad \alpha > 1.$$

In particular, we have

$$\frac{(\log n)^\alpha}{2} < n2^{-j} \leq (\log n)^\alpha.$$

Now,

$$\hat\beta_{jk} = \frac{1}{n}\sum_{i=1}^n \psi_{jk}(X_i) = \frac{2^{\frac{j}{2}}}{n}(N_{jk}^+ - N_{jk}^-).$$

Hence,

$$\begin{aligned} \mathbb{E}(\|\tilde f_{n,\gamma} &- f\|_2^2) \geq \sum_{k=0}^{2^j-1} \mathbb{E}\left(\hat\beta_{jk}^2 \mathbb{1}_{|\hat\beta_{jk}|\geq \eta_{jk,\gamma}}\right) \\ &\geq \sum_{k=0}^{2^j-1} \mathbb{E}\left(\hat\beta_{jk}^2 \mathbb{1}_{|\hat\beta_{jk}|\geq \widetilde{\eta_{jk,\gamma}}} \mathbb{1}_{|u_n|\leq U(2\log n)}\right) \\ &\geq \sum_{k=0}^{2^j-1} \frac{2^j}{n^2}\mathbb{E}\left((N_{jk}^+ - N_{jk}^-)^2 \mathbb{1}_{|\hat\beta_{jk}|\geq \sqrt{2\gamma(1+\varepsilon)s_n\frac{\log n}{n}} + w(\varepsilon)\frac{2^{j/2}\log n}{n}} \mathbb{1}_{|u_n|\leq U(2\log n)}\right). \\ &\geq \sum_{k=0}^{2^j-1} \frac{2^j}{n^2}\mathbb{E}\left((N_{jk}^+ - N_{jk}^-)^2 \mathbb{1}_{\frac{2^{\frac{j}{2}}}{n}|N_{jk}^+ - N_{jk}^-|\geq \sqrt{2\gamma(1+\varepsilon)\frac{2^j}{n}\left(N_{jk}^+ + N_{jk}^-\right)\frac{\log n}{n}} + w(\varepsilon)\frac{2^{j/2}\log n}{n}} \mathbb{1}_{|u_n|\leq U(2\log n)}\right) \\ &\geq \sum_{k=0}^{2^j-1} \frac{2^j}{n^2}\mathbb{E}\left((N_{jk}^+ - N_{jk}^-)^2 \mathbb{1}_{|N_{jk}^+ - N_{jk}^-|\geq \sqrt{2\gamma(1+\varepsilon)\left(N_{jk}^+ + N_{jk}^-\right)\log n} + w(\varepsilon)\log n} \mathbb{1}_{|u_n|\leq U(2\log n)}\right) \\ &\geq \frac{2^{2j}}{n^2}\mathbb{E}\left((N_{j1}^+ - N_{j1}^-)^2 \mathbb{1}_{|N_{j1}^+ - N_{j1}^-|\geq \sqrt{2\gamma(1+\varepsilon)\left(N_{j1}^+ + N_{j1}^-\right)\log n} + w(\varepsilon)\log n} \mathbb{1}_{|u_n|\leq U(2\log n)}\right). \end{aligned}$$

Now, we consider a bounded sequence $(w_n)_n$ such that for any n, $w_n \geq w(\varepsilon)$ and such that $\frac{\sqrt{v_{nj}}}{2}$ is an integer with

$$v_{nj} = \left(\sqrt{4\gamma(1+\varepsilon)\tilde\mu_{nj}\log(n)} + w_n\log(n)\right)^2$$

and $\tilde\mu_{nj}$ is the largest integer smaller or equal to $n2^{-j-1}$. We have

$$v_{nj} \sim 4\gamma(1+\varepsilon)\tilde\mu_{nj}\log n$$

and

$$\frac{(\log n)^\alpha}{4} - 1 < n2^{-j-1} - 1 < \tilde\mu_{nj} \leq n2^{-j-1} \leq \frac{(\log n)^\alpha}{2}.$$

So, if

$$N^+_{j1} = \tilde{\mu}_{nj} + \frac{1}{2}\sqrt{v_{nj}}, \quad N^-_{j1} = \tilde{\mu}_{nj} - \frac{1}{2}\sqrt{v_{nj}},$$

then

$$N^+_{j1} + N^-_{j1} = 2\tilde{\mu}_{nj}, \quad N^+_{j1} - N^-_{j1} = \sqrt{v_{nj}} = \sqrt{2\gamma(1+\varepsilon)\left(N^+_{j1} + N^-_{j1}\right)\log n} + w_n \log n.$$

Finally,

$$\begin{aligned}
\mathbb{E}(\|\tilde{f}_{n,\gamma} - f\|_2^2) &\geq \frac{2^{2j}}{n^2} v_{nj} \mathbb{P}\left(N^+_{j1} = \tilde{\mu}_{nj} + \frac{1}{2}\sqrt{v_{nj}}, \quad N^-_{j1} = \tilde{\mu}_{nj} - \frac{1}{2}\sqrt{v_{nj}}, \quad |u_n| \leq U(2\log n)\right) \\
&\geq v_{nj}(\log n)^{-2\alpha} \\
&\quad \times \left[\mathbb{P}\left(N^+_{j1} = \tilde{\mu}_{nj} + \frac{1}{2}\sqrt{v_{nj}}, \quad N^-_{j1} = \tilde{\mu}_{nj} - \frac{1}{2}\sqrt{v_{nj}}\right) - \mathbb{P}\left(|u_n| > U(2\log n)\right)\right] \\
&\geq v_{nj}(\log n)^{-2\alpha}\left[\frac{n!}{l_{nj}!m_{nj}!(n - l_{nj} - m_{nj})!} p_j^{l_{nj}+m_{nj}} (1-2p_j)^{n-(l_{nj}+m_{nj})} - \frac{6}{n^2}\right],
\end{aligned}$$

with

$$l_{nj} = \tilde{\mu}_{nj} + \frac{1}{2}\sqrt{v_{nj}}, \quad m_{nj} = \tilde{\mu}_{nj} - \frac{1}{2}\sqrt{v_{nj}},$$

and

$$p_j = \int \mathbb{1}_{[k2^{-j},(k+0.5)2^{-j}[}(x) f(x) dx = \int \mathbb{1}_{[(k+0.5)2^{-j},(k+1)2^{-j}[}(x) f(x) dx = 2^{-j-1}.$$

So,

$$\mathbb{E}(\|\tilde{f}_{n,\gamma} - f\|_2^2) \quad \geq \quad v_{nj}(\log n)^{-2\alpha} \times \left[\frac{n!}{l_{nj}!m_{nj}!(n - 2\tilde{\mu}_{nj})!} p_j^{2\tilde{\mu}_{nj}} (1-2p_j)^{n-2\tilde{\mu}_{nj}} - \frac{6}{n^2}\right].$$

Now, let us study each term:

$$\begin{aligned}
p_j^{2\tilde{\mu}_{nj}} &= \exp\left(2\tilde{\mu}_{nj}\log(p_j)\right) \\
&= \exp\left(2\tilde{\mu}_{nj}\log(2^{-j-1})\right),
\end{aligned}$$

$$\begin{aligned}
(1-2p_j)^{n-2\tilde{\mu}_{nj}} &= \exp\left((n - 2\tilde{\mu}_{nj})\log(1-2p_j)\right) \\
&= \exp\left(-(n - 2\tilde{\mu}_{nj})\left(2^{-j} + O_n(2^{-2j})\right)\right) \\
&= \exp\left(-n2^{-j}\right)(1 + o_n(1)),
\end{aligned}$$

$$n! \quad = \quad n^n e^{-n}\sqrt{2\pi n}\,(1 + o_n(1)),$$

$$\begin{aligned}
(n - 2\tilde{\mu}_{nj})^{n-2\tilde{\mu}_{nj}} &= \exp\left(\left(n - 2\tilde{\mu}_{nj}\right)\log\left(n - 2\tilde{\mu}_{nj}\right)\right) \\
&= \exp\left(\left(n - 2\tilde{\mu}_{nj}\right)\left(\log n + \log\left(1 - \frac{2\tilde{\mu}_{nj}}{n}\right)\right)\right) \\
&= \exp\left(\left(n - 2\tilde{\mu}_{nj}\right)\log n - \frac{2\tilde{\mu}_{nj}\left(n - 2\tilde{\mu}_{nj}\right)}{n}\right)(1 + o_n(1)) \\
&= \exp\left(n\log n - 2\tilde{\mu}_{nj} - 2\tilde{\mu}_{nj}\log n\right)(1 + o_n(1)).
\end{aligned}$$

Then,

$$\begin{aligned}
\frac{n!}{(n-2\tilde{\mu}_{nj})!}p_j^{2\tilde{\mu}_{nj}}(1-2p_j)^{n-2\tilde{\mu}_{nj}} &= \frac{e^{n-2\tilde{\mu}_{nj}}}{e^n}\times\frac{n^n}{(n-2\tilde{\mu}_{nj})^{n-2\tilde{\mu}_{nj}}}\times p_j^{2\tilde{\mu}_{nj}}(1-2p_j)^{n-2\tilde{\mu}_{nj}}(1+o_n(1))\\
&= \exp\left(-2\tilde{\mu}_{nj}\right)\times\frac{\exp(n\log n)}{(n-2\tilde{\mu}_{nj})^{n-2\tilde{\mu}_{nj}}}\times p_j^{2\tilde{\mu}_{nj}}(1-2p_j)^{n-2\tilde{\mu}_{nj}}(1+o_n(1))\\
&= \exp\left(-2\tilde{\mu}_{nj}\right)\times\frac{\exp\left(n\log n+2\tilde{\mu}_{nj}\log(2^{-j-1})-n2^{-j}\right)}{\exp\left(n\log n-2\tilde{\mu}_{nj}-2\tilde{\mu}_{nj}\log n\right)}(1+o_n(1))\\
&= \exp\left(2\tilde{\mu}_{nj}\log n+2\tilde{\mu}_{nj}\log(2^{-j-1})-n2^{-j}\right)(1+o_n(1)).
\end{aligned}$$

It remains to evaluate $l_{nj}!\times m_{nj}!$

$$\begin{aligned}
l_{nj}!\times m_{nj}! &= \left(\frac{l_{nj}}{e}\right)^{l_{nj}}\left(\frac{m_{nj}}{e}\right)^{m_{nj}}\sqrt{2\pi l_{nj}}\sqrt{2\pi m_{nj}}(1+o_n(1))\\
&= \exp\left(l_{nj}\log l_{nj}+m_{nj}\log m_{nj}-2\tilde{\mu}_{nj}\right)\times 2\pi\tilde{\mu}_{nj}(1+o_n(1)).
\end{aligned}$$

If we set

$$x_{nj}=\frac{\sqrt{\nu_{nj}}}{2\tilde{\mu}_{nj}}=o_n(1),$$

then

$$l_{nj}=\tilde{\mu}_{nj}+\frac{\sqrt{\nu_{nj}}}{2}=\tilde{\mu}_{nj}(1+x_{nj}),$$

$$m_{nj}=\tilde{\mu}_{nj}-\frac{\sqrt{\nu_{nj}}}{2}=\tilde{\mu}_{nj}(1-x_{nj}),$$

and using that

$$\begin{aligned}
(1+x_{nj})\log(1+x_{nj}) &= (1+x_{nj})\left(x_{nj}-\frac{x_{nj}^2}{2}+\frac{x_{nj}^3}{3}+O(x_{nj}^4)\right)\\
&= x_{nj}-\frac{x_{nj}^2}{2}+\frac{x_{nj}^3}{3}+x_{nj}^2-\frac{x_{nj}^3}{2}+O(x_{nj}^4)\\
&= x_{nj}+\frac{x_{nj}^2}{2}-\frac{x_{nj}^3}{6}+O(x_{nj}^4)
\end{aligned}$$

$$\begin{aligned}
l_{nj}\log l_{nj} &= \tilde{\mu}_{nj}(1+x_{nj})\log\left(\tilde{\mu}_{nj}(1+x_{nj})\right)\\
&= \tilde{\mu}_{nj}(1+x_{nj})\log(1+x_{nj})+\tilde{\mu}_{nj}(1+x_{nj})\log\left(\tilde{\mu}_{nj}\right)\\
&= \tilde{\mu}_{nj}\left(x_{nj}+\frac{x_{nj}^2}{2}-\frac{x_{nj}^3}{6}+O(x_{nj}^4)\right)+\tilde{\mu}_{nj}(1+x_{nj})\log\left(\tilde{\mu}_{nj}\right).
\end{aligned}$$

Similarly,

$$m_{nj}\log m_{nj} = \tilde{\mu}_{nj}\left(-x_{nj}+\frac{x_{nj}^2}{2}+\frac{x_{nj}^3}{6}+O(x_{nj}^4)\right)+\tilde{\mu}_{nj}(1-x_{nj})\log\left(\tilde{\mu}_{nj}\right).$$

So,

$$\begin{aligned}
l_{nj}\log l_{nj}+m_{nj}\log m_{nj} &= \tilde{\mu}_{nj}\left(x_{nj}^2+O(x_{nj}^4)\right)+2\tilde{\mu}_{nj}\log\left(\tilde{\mu}_{nj}\right)\\
&\le \tilde{\mu}_{nj}x_{nj}^2+2\tilde{\mu}_{nj}\log(n2^{-j-1})+O(\tilde{\mu}_{nj}x_{nj}^4).
\end{aligned}$$

Since

$$\tilde{\mu}_{nj} x_{nj}^2 = \frac{v_{nj}}{4\tilde{\mu}_{nj}} \sim \gamma(1+\varepsilon)\log n,$$

for n large enough,

$$\tilde{\mu}_{nj} x_{nj}^2 + O(\tilde{\mu}_{nj} x_{nj}^4) \leq (\gamma + 2\varepsilon)\log n$$

and

$$l_{nj}\log l_{nj} + m_{nj}\log m_{nj} \leq (\gamma + 2\varepsilon)\log n + 2\tilde{\mu}_{nj}\log(n2^{-j-1}).$$

Finally,

$$\begin{aligned} l_{nj}! \times m_{nj}! &= \exp\left(l_{nj}\log l_{nj} + m_{nj}\log m_{nj} - 2\tilde{\mu}_{nj}\right) 2\pi\tilde{\mu}_{nj}(1+o_n(1)) \\ &\leq \exp\left((\gamma+2\varepsilon)\log n + 2\tilde{\mu}_{nj}\log(n2^{-j-1}) - 2\tilde{\mu}_{nj}\right) 2\pi\tilde{\mu}_{nj}(1+o_n(1)). \end{aligned}$$

we derive that

$$\begin{aligned} \mathbb{E}(\|\tilde{f}_{n,\gamma} &- f\|_2^2) \geq v_{nj}(\log n)^{-2\alpha}\left[\frac{n!}{l_{nj}!m_{nj}!(n-2\tilde{\mu}_{nj})!} p_j^{2\tilde{\mu}_{nj}}(1-2p_j)^{n-2\tilde{\mu}_{nj}} - \frac{6}{n^2}\right] \\ &\geq v_{nj}(\log n)^{-2\alpha}\left[\frac{\exp\left(2\tilde{\mu}_{nj}\log n + 2\tilde{\mu}_{nj}\log(2^{-j-1}) - n2^{-j}\right)}{\exp\left((\gamma+2\varepsilon)\log n + 2\tilde{\mu}_{nj}\log(n2^{-j-1}) - 2\tilde{\mu}_{nj}\right)\times 2\pi\tilde{\mu}_{nj}} - \frac{6}{n^2}\right](1+o_n(1)) \\ &\geq v_{nj}(\log n)^{-2\alpha}\left[\frac{\exp\left(-(\gamma+2\varepsilon)\log n - 2\right)}{2\pi\tilde{\mu}_{nj}} - \frac{6}{n^2}\right](1+o_n(1)) \end{aligned}$$

So there exists C_1 and C_2 two positive constants such that, for n large enough

$$\mathbb{E}(\|\tilde{f}_{n,\gamma} - f\|_2^2) \geq C_1(\log n)^{1-\alpha}\left[C_2\frac{n^{-(\gamma+2\varepsilon)}}{(\log n)^{\alpha}} - \frac{6}{n^2}\right].$$

As $0 < \gamma + 2\varepsilon < 1$, there exists a positive constant $\delta < 1$ such that

$$\mathbb{E}(\|\tilde{f}_{n,\gamma} - f\|_2^2) \geq \frac{1}{n^{\delta}}(1+o_n(1)).$$

This concludes the proof of Theorem 2.

Ackowledgment: The authors acknowledge Rebecca Willett for her very smart program, and both A. Antoniadis and L. Birgé for a wealth of advice and encouragement. We also thank the anonymous referees for helpful comments and valuable suggestions.

References

Antoniadis, A., Grégoire, G. and Nason, G. (1999). Density and hazard rate estimation for right censored data using wavelet methods. *Journ. Royal Statist. Soc. B*, **61**(1), 63–84.

Bertin, K., Le Pennec, E. and Rivoirard, V. (2010). Adaptive Dantzig density estimation. *Annales de l'Institut Henri Poincaré*, to appear.

Birgé, L. (2008). Model selection for density estimation with L_2-loss. Technical report. http://hal.archives-ouvertes.fr/hal-00347691_v1.

Birgé, L. and Massart, P. (2007). Minimal penalties for Gaussian model selection. *Probab. Theory Related Fields*, **138**(1-2), 33–73.

Birgé, L. and Rozenholc, Y. (2006). How many bins should be put in a regular histogram. *ESAIM Probab. Stat.* , **10**, 24–45.

Brown, L., Cai, T., Zhang, R., Zhao, L. and Zhou, H. (2010). The root-unroot algorithm for density estimation as implemented via wavelet block thresholding. *Probability Theory and Related Fields*, **146**(3-4), 401–433.

Castellan, G. (2000). Sélection d'histogrammes à l'aide d'un critère de type Akaike. *C. R. Acad. Sci. Paris Sr. I Math.*, **330**(8), 729–732.

Castellan, G. (2003). Density estimation via exponential model selection. *IEEE Trans. Inform. Theory*, **49**(8), 2052–2060.

Cohen, A., Daubechies, I. and Feauveau, J.C. (1992). Biorthogonal bases of compactly supported wavelets. *Comm. Pure Appl. Math.*, **45**(5), 485–560.

Copas, J.B. and Fryer, M.J. (1980). Density estimation and suicide risks in psychiatric treatment. *J. Roy. Statist. Soc. A*, **143**, 167–176.

Delyon, B. and Juditsky, A. (1997). On the computation of wavelet coefficients. *J. Approx. Theory*, **88**(1), 47–79.

DeVore, R.A. and Lorentz, G.G. (1993). *Constructive approximation*. Berlin: Springer-Verlag.

Donoho, D.L. (1994). Smooth wavelet decompositions with blocky coefficient kernels. *Recent advances in wavelet analysis, Wavelet Anal. Appl.* **3** Academic Press, Boston, MA 259–308.

Donoho, D.L. and Johnstone, I.M. (1994). Ideal spatial adaptation by wavelet shrinkage. *Biometrika*, **81**(3), 425–455.

Donoho, D.L., Johnstone, I.M., Kerkyacharian, G. and Picard, D. (1996). Density estimation by wavelet thresholding. *Annals of Statistics*, **24**(2), 508–539.

Giné, E., Latala, R. and Zinn, J. (2000). Exponential and moment inequalities for U-statistics. *High Dimensional Probability II, Progr. Probab.*, **47**, Birkhäuser, Boston. 13–38.

Giné, E. and Nickl, R. (2009). Uniform limit theorems for wavelet density estimators. *Annals of Probability*, **37**(4), 1605–1646.

Giné, E. and Nickl, R. (2010). Adaptive estimation of the distribution function and its density in sup-norm loss by wavelet and spline projections. *Bernoulli*, to appear.

Härdle, W., Kerkyacharian, G., Picard, D. and Tsybakov, A. (1998). *Wavelets, approximation and statistical applications*. Lecture Notes in Statistics, **129**, New York: Springer-Verlag.

Herrick, D. R. M., Nason, G. P. and Silverman, B. W. (2001). Some new methods for wavelet density estimation. *Sankhya Ser. A*, **63**, (3) 394–411.

Houdré, C. and Reynaud-Bouret, P. (2003). Exponential inequalities, with constants, for U-statistics of order two. *Stochastic inequalities and applications, Progr. Probab.*, **56**, Birkhäuser, Basel. 55–69.

Johnstone, I.M. (1994). *Minimax Bayes, asymptotic minimax and sparse wavelet priors*. Statistical decision theory and related topics, V (West Lafayette, IN, 1992), 303–326. New York: Springer.

Juditsky, A. and Lambert-Lacroix, S. (2004). On minimax density estimation on $\mathbb{R}$. *Bernoulli*, **10**(2), 187–220.

Klein, T. and Rio, E. (2005). Concentration around the mean for maxima of empirical processes. *Ann. Proba.*, **33**(3), 1060–1077.

Koo, J-Y., Kooperberg, C. and Park, J. (1999). Logspline density estimation under censoring and truncation. *Scand. J. Statist.*, **26**(1), 87–105.

Lebarbier, E. (2005). Detecting multiple change-points in the mean of Gaussian process by model selection. *Signal Processing*, **85**(4), 717–736.

Massart, P. (2007). *Concentration inequalities and model selection.* Lectures from the 33rd Summer School on Probability Theory held in Saint-Flour, July 6–23, 2003. Berlin: Springer.

Reynaud-Bouret, P. and Rivoirard, V. (2010). Near optimal thresholding estimation of a Poisson intensity on the real line. *EJS*, **4**, 172–238.

Rudemo, M. (1982). Empirical choice of histograms and density estimators. *Scand. J. Statist.*, **9**(2), 65–78.

Sain, S.R. and Scott, D.W. (1996). On locally adaptive density estimation. *J. Amer. Statist. Assoc.* , **91**(436), 1525–1534.

Shorack, G. R. and Wellner, J. A. (1986). *Empirical processes with applications to statistics*. Wiley Series in Probability and Mathematical Statistics: Probability and Mathematical Statistics. New York: John Wiley & Sons, Inc..

Silverman, B.W. (1986). *Density Estimation for Statistics and Data Analysis.* Monograph on Statistics and Applied Probability, **26**. Chapman & Hall.

Weisberg, S. (1980). *Applied Linear Regression*. New-York: Wiley.

Willett, R.M. and Nowak, R.D. (2007). Multiscale Poisson Intensity and Density Estimation. *IEEE Transactions on Information Theory*, **53**(9), 3171–3187.